

멀티미디어 데이터베이스 시스템에서 최대거리를 이용한 K-최대근접질의 처리 방법 (The Method to Process Nearest Neighbor Queries using Maximun Distance in Multimedia Database Systems)

선희준(Hwi-Joon Seon)¹⁾ 신성철(Seong-Chul Shin)²⁾

요 약

멀티미디어 데이터베이스 시스템에서 k-최대근접질의는 매우 빈번히 발생하며, 다른 유형의 공간질의에 비하여 처리비용이 많이 요구된다. K-최대근접질의의 처리비용을 최적화하기 위해서는 색인에서 검색되는 노드의 수와 연산시간을 최소화할 수 있어야 한다. 본 논문에서는 최적검색거리[1]의 연산시간을 줄일 수 있는 새로운 검색거리를 제안하고 그 특성을 보인다.

ABSTRACT

In multimedia database systems, the k nearest neighbor query occurs frequently and requires the processing cost higher than other spatial queries do. The number of searched nodes and the computation time in an index can be minimized for optimizing the cost of processing the k nearest neighbor query. In this paper, we propose the search distance which can reduce the computation time of the optimal search distance.

논문접수 : 2004. 11. 30.

심사완료 : 2004. 12. 15.

1) 정희원 :서남대학교 컴퓨터정보통신학과

2) 정희원 : 해군사관학교 전산학과

1. 서론

최근의 정보 서비스들은 대용량의 멀티미디어 데이터베이스 시스템을 기반으로 하고 있다. 이러한 멀티미디어 데이터베이스 시스템에서는 주어진 위치에서 가장 근접한 객체를 찾는 k-최대 근접 질의가 자주 발생한다 [3,4,5,6,7,8]. 그러나 k-최대근접질의의 처리는 연산 및 보조기억장치 접근을 위한 많은 처리 시간이 요구된다.

최대 근접 질의의 처리비용을 최적화하기 위해서는 연산 시간과 색인에서 검색되는 노드의 수를 최소화할 수 있어야 한다. 이를 위해 최대 근접 질의의 처리시 색인에서 방문될 노드들이 정확히 선정되도록 위치 속성에 의한 검색거리 측도인 최적검색거리가 제안되었다 [1]. 최적검색거리는 질의기준으로부터 객체 또는 부검색공간들이 반드시 존재하는 거리 중에서 최소의 거리이며, 최대 근접 질의의 처리시 질의기준의 유형에 관계없이 색인에서 검색될 노드들을 정확히 선택하기 위한 검색거리 측도이다.

본 논문에서는 대용량의 멀티미디어 데이터베이스 환경에서 R-트리 유형의 색인구조를 이용한 k-최대 근접 질의 처리방법에 따른 연산시간을 최소화할 수 있는 새로운 검색거리를 제안한다.

2. 관련연구

최적검색거리를 이용한 방법[1]에서는 최대 근접 질의 처리를 위한 기본개념과 기존의 측정방법이 가지고 있는 문제점을 해결한 새로운 검색거리 측도인 확장최소거리 및 최적검색거리를 정의하였다. N차원 검색공간에서 질의기준 Q와 최소경계사각형(minimum bounding rectangle: M)사이의 가장 가까운 거리인 확장최소거리(eXtended MINimum DISTance: XMINDIST)는 다음과 같다.

$$XMINDIST(Q, M) = \sum_{i=1}^N |Q_i - M_i|^2$$

여기에서

$$Q_i = \begin{cases} Q_{Li}, & Q_{Li} > M_{Li} \\ Q_{Ui}, & Q_{Li} < M_{Li} \\ 0, & otherwise \end{cases} \quad M_i = \begin{cases} M_{Li}, & Q_{Li} < M_{Li} \\ M_{Ui}, & Q_{Li} > M_{Li} \\ 0, & otherwise. \end{cases}$$

(Q_{Li}, Q_{Ui} : 질의범위의 시작과 끝,
 M_{Li}, M_{Ui} : 최소경계사각형의 시작과 끝)

정의된 XMINDIST는 최소경계사각형 M에 포함되어 있는 부검색공간들 중에서 질의기준 Q에 가장 근접하고 있는 객체 또는 부검색공간을 결정하기 위한 거리이다.

최적검색거리(the Optimized MINimum value of all DISTances: OMINDIST)는 N차원 검색공간상에서 질의기준 Q로부터 최소경계사각형 M을 구성하는 임의의 N-1차원을 포함할 수 있는 거리들 중 최소거리이며 다음과 같다.

만약 Q와 M이 겹쳐있으면

$$OMINDIST(Q, M) = \min_{1 \leq i \leq N} \left\{ \begin{aligned} & (|qr_i - mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - Mr_k|^2), \\ & (|qr_i - Mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - mr_k|^2), \\ & (|q_i - Mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Q_k - mr_k|^2) \end{aligned} \right\}$$

여기에서

$$qr_i = \begin{cases} Q_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \geq \frac{(M_{Li} + M_{Ui})}{2} \\ Q_{Ui}, & otherwise \end{cases}$$

$$Qr_k = \begin{cases} Q_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \leq \frac{(M_{Lk} + M_{Uk})}{2} \\ Q_{Uk}, & otherwise \end{cases}$$

$$q_i = \begin{cases} Q_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \geq \frac{(M_{Li} + M_{Ui})}{2} \\ Q_{Ui}, & otherwise \end{cases}$$

$$Q_k = \begin{cases} Q_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ Q_{Uk}, & \text{otherwise} \end{cases}$$

$$mr_i = \begin{cases} M_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \leq \frac{(M_{Li} + M_{Ui})}{2} \\ M_{Ui}, & \text{otherwise} \end{cases}$$

$$Mr_k = \begin{cases} M_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ M_{Uk}, & \text{otherwise.} \end{cases}$$

그렇지 않으면

$$OMINDIST(Q, M) = \min_{1 \leq i \leq N} \left\{ (|qr_i - mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - Mr_k|^2), \right. \\ \left. (|qr_i - Mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - mr_k|^2) \right\}$$

여기에서

$$qr_i = \begin{cases} Q_{Li}, & Q_{Li} \geq M_{Ui} \\ Q_{Ui}, & Q_{Ui} \leq M_{Li} \\ Q_{ii}, & \text{otherwise} \end{cases}$$

$$Q_i = \begin{cases} Q_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \leq \frac{(M_{Li} + M_{Ui})}{2} \\ Q_{Ui}, & \text{otherwise} \end{cases}$$

$$Qr_k = \begin{cases} Q_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ Q_{Uk}, & \text{otherwise} \end{cases}$$

$$mr_i = \begin{cases} M_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \leq \frac{(M_{Li} + M_{Ui})}{2} \\ M_{Ui}, & \text{otherwise} \end{cases}$$

$$Mr_k = \begin{cases} M_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ M_{Uk}, & \text{otherwise} \end{cases}$$

OMINDIST는 질의기준 Q로부터 최소경계 사각형 M에 따른 부검색공간에 포함된 객체 또는 부검색공간들이 적어도 하나는 반드시 존재하는 최적의 거리이다.

최적검색거리를 이용한 방법[1]에서는 OMINDIST를 적용함으로써 질의 처리시 검색되어야 하는 노드를 정확히 선택할 수 있음을

정리하였으며, 이는 색인에서 방문되는 전체 노드의 수가 최소화됨을 나타내기 때문에 질의 처리에 따른 비용을 최적화함을 보였다.

3. 개선된 k-최대근접질의 처리 방법

기존의 k-최대근접질의 처리 방법은 확장최소거리 XMINDIST와 최적검색거리 OMINDIST를 조합하여 검색 대상이 되는 노드의 수를 최소화하였다. 그러나 OMINDIST의 연산 시간을 최적화하지 못했다. 본 논문에서는 검색대상이 되는 노드의 선정 시 연산 시간을 최소화할 수 있는 최대근접질의 처리 방법을 제안한다.

색인을 이용하여 주어진 질의 기준으로부터 가장 가까운 객체를 검색하는 동안 XMINDIST와 OMINDIST의 차이가 아주 크거나 작은 경우가 발생할 수 있다. 이를 위해 질의기준과 최소경계사각형의 모든 꼭지점들 중에서 가장 먼 점까지의 거리인 최대거리(the MAXimum DISTance value of all distances: MAXDIST)를 사용한다. MAXDIST를 사용함으로써 k-최대 근접 질의 처리시 색인에서 방문할 노드의 수와 연산시간을 최소화할 수 있다.

【정의 1】 N차원 검색공간에서 질의기준 Q와 최소경계사각형 M간의 최대거리 MAXDIST를 다음과 같이 정의한다.

$$\max_{1 \leq i \leq N} \left\{ (|qr_i - Mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - Mr_k|^2), \right. \\ \left. (|qr_i - Mr_i|^2 + \sum_{\substack{i \neq k \\ 1 \leq k \leq N}} |Qr_k - mr_k|^2) \right\}$$

여기에서

$$qr_i = \begin{cases} Q_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \geq \frac{(M_{Li} + M_{Ui})}{2} \\ Q_{Ui}, & \text{otherwise} \end{cases}$$

$$Qr_k = \begin{cases} Q_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \leq \frac{(M_{Lk} + M_{Uk})}{2} \\ Q_{Uk}, & \text{otherwise} \end{cases}$$

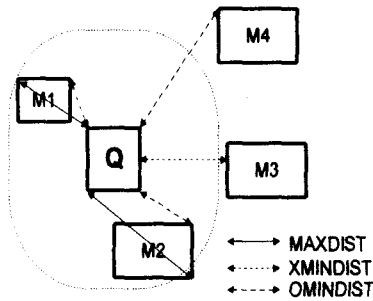
$$q_i = \begin{cases} Q_{Li}, & \text{if } \frac{(Q_{Li} + Q_{Ui})}{2} \geq \frac{(M_{Li} + M_{Ui})}{2} \\ Q_{Ui}, & \text{otherwise} \end{cases}$$

$$Q_k = \begin{cases} Q_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ Q_{Uk}, & \text{otherwise} \end{cases}$$

$$Mr_k = \begin{cases} M_{Lk}, & \text{if } \frac{(Q_{Lk} + Q_{Uk})}{2} \geq \frac{(M_{Lk} + M_{Uk})}{2} \\ M_{Uk}, & \text{otherwise.} \end{cases}$$

□

그림 1은 질의 기준 Q로부터 XMINDIST가 가장 작은 최소경계사각형 M1에 대한 MAXDIST를 기준으로 다음 방문대상에서 노드들을 제외하는 예이다. k개의 최대 근접 객



[그림 1] 최대거리 MAXDIST의 예
[Fig. 1] Example of MAXDIST

체의 검색시 MAXDIST의 사용은 하나의 최소 경계사각형에 포함되는 모든 객체 또는 부검색 공간들과 비교가 가능하므로 색인에서 방문되는 노드의 수와 연산시간을 최소화할 수 있다.

정의된 MAXDIST는 OMINDIST보다 작은

연산 시간을 갖으며 하위 레벨의 노드에 있는 엔트리의 수를 고려한 새로운 검색거리이다. k-최대근접질의 알고리즘에서는 R-트리를 검색하는 동안 방문할 필요가 없는 노드들을 방문대상에서 제외하기 위해 OMINDIST 대신에 MAXDIST만을 이용하여 다음과 같은 전략을 사용한다.

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

ii) 질의 기준 Q로부터 현재까지 발견된 k번째로 가까운 객체 O의 거리가 최소경계사각형 M에 대한 MAXDIST(Q,M)보다 크다면, 객체 O를 최대근접객체 대상에서 제외한다.

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

4. 최대거리 특성

다음의 정리들은 제안된 MAXDIST을 이용한 방법과 OMINDIST을 이용한 k-최대근접질의 처리 방법과의 특성을 기술한 것이다.

【정리1】 P가 노드 A에 포함되어 있는 객체이고 질의 Q까지 가장 근접한 객체라고 하면 $XMINDIST_A \leq DIST_P \leq MAXDIST_A$ 이다.

(증명) 확장최소거리 정의에 의해 $XMINDIST_A$ 는 A에서 질의 Q까지 최소 거리이다. 그리고 P를 A에 포함되어 있는 객체들 중에서 질의로부터 가장 가까운 객체라고 하면, $DIST_P \leq MAXDIST_A$ 가 성립한다. 그러므로 $XMINDIST_A$ 와

$MAXDIST_A$ 는 노드 A 에 존재하는 최대근접객체에서 질의까지의 최소거리와 최대거리가 된다. 따라서 $XMINDIST_A \leq DIST_P \leq MAXDIST_A$ 가 성립한다.

【정리2】 A 가 B 의 상위 노드이면 $XMINDIST_A \leq XMINDIST_B$ 이다.

(증명) 확장최소거리 $XMINDIST$ 의 정의에 의해 $XMINDIST_A \leq XMINDIST_B$ 가 성립한다.

【정리3】 노드 A 와 B 가 $XMINDIST_A \leq XMINDIST_B$ 라고 했을 때 $OMINDIST_B < XMINDIST_A$ 는 성립하지 않는다.

(증명) 보조정리 1에 의해 $XMINDIST_B \leq OMINDIST_B$ 는 성립한다.

$XMINDIST_A \leq XMINDIST_B$ 이므로 $XMINDIST_A \leq XMINDIST_B \leq OMINDIST_B$ 이다. 따라서 $OMINDIST_B < XMINDIST_A$ 는 성립하지 않는다.

【정리4】 $XMINDIST_A > OMINDIST_B$ 인 경우 노드 A 가 방문대상에서 제거된다면 $XMINDIST_A > MAXDIST_B$ 인 경우에도 제거된다.

(증명)

(1) $XMINDIST_A \leq XMINDIST_B$ 인 경우 정리 1에 의해 $XMINDIST_A \leq OMINDIST_A$ 와 $XMINDIST_B \leq OMINDIST_B$ 는 성립하지 않는다. $XMINDIST_A \leq XMINDIST_B$ 이므로 $OMINDIST_B < XMINDIST_A$ 는 성

립하지 않는다. 따라서 노드 A 는 방문대상에서 제거된다.

(2) $XMINDIST_A > XMINDIST_B$ 이고 노드 A 가 노드 B 이전에 검색되는 경우

$XMINDIST$ 에 의해 검색 순서를 결정하는 것은 가능하지 않다.

(3) $XMINDIST_A > XMINDIST_B$ 이고 노드 B 가 노드 A 이전에 검색되는 경우

$MAXDIST_{temp}$ 을 A 가 검색대상이 되거나 제거대상이 되기 전 상태의 최대거리라고 하면 $MAXDIST_{temp} < OMINDIST_A$ 이다. 그러므로 $OMINDIST_B < XMINDIST_A$ 이면 $MAXDIST_{temp} < XMINDIST_A$ 가 된다. 따라서 노드 A 는 검색대상에서 제거된다.

5. 결론

대용량의 멀티미디어 데이터베이스 시스템에서 k 개의 최대 근접 객체를 찾는 질의의 처리는 많은 디스크 접근과 질의처리시간을 요구한다. 또한 데이터의 차원이 증가함에 따라 검색비용이 크게 증가할 수 있다.

본 논문에서는 멀티미디어 데이터베이스 시스템에서 k -최대 근접 질의의 처리시 연산시간을 최소화하기 위한 최대거리를 제안하고 그 특성을 정리하였다. 제안된 알고리즘은 색인에서 하위 노드의 엔트리 수가 고려된 최대거리를 이용함으로써 k -최대근접질의 처리시 기존의 방법보다 더 작은 질의처리시간을 요구한다.

앞으로의 연구방향은 최대거리를 이용한 질의처리 방법과 최적검색거리를 이용한 질의처리 방법의 성능을 실험을 통하여 비교 평가하는 것이다.

참고문헌

- [1] 선휘준 외 1, "최적탐색거리를 이용한 최근접질의 처리방법의 성능 평가," 한국정보처리학회 논문지, 6권 1호, pp. 32-41, 1999.1.
- [2] N.Beckmann,H.Kriegel, R.Schneider and B.Seeger,"The R*-tree:an Efficient and Robust Access Method for Points and Rectangles," Proc. ACM SIGMOD Int. Conf. on Management of Data, pp. 322-331, 1990
- [3]C. Faloutsos. et al., "Efficient and Effective Querying by Image Content," Journal of Intelligent Information Systems, Vol. 3, No. 4, pp.231-262, 1994.
- [4] D.A.White, R.Jain.,"Similarity Indexing with SS-tree," In Proc. Intl. Conf. on Data Engineering, pp.516-523, 1991.
- [5]M.Flinker, et al.,"Query by Image and Video Content: The QBIC Sytem," IEEE Computer, Vol.28, No.9, pp.23-32, 1995.
- [6]N.Roussopoulos, et al., "Nearest Neighbor Queries," In Proc. Intl. Conf. on Management of Data, ACM SIGMOD, pp.71-79, 1995.
- [7]S.A.Nene, S.K.Nayar, "Closest Point Search in High Dimensions," In Proc. Intl. Conf. on Computer Vision and Pattern Recognition, 1996.
- [8]S.Berchtold,et al,"Fast Parallel Similarity Search in Multimedia Databases," In Proc. of ACM SIGMOD Int. Conf. on the Management of Data, 1997.