

# 도로 네트워크 데이터베이스를 위한 근사 인덱싱

이상철, 김상욱  
한양대학교 전자컴퓨터통신공학과  
e-mail: korly@zion.hanyang.ac.kr

## Approximate Indexing in Road Network Databases

Sang-Chul Lee, Sang-Wook Kim  
Department of Electronics and Computer Engineering, Hanyang University

### 요 약

본 논문에서는 도로 네트워크 데이터베이스에서 k-최근접 이웃 질의를 효율적으로 처리하기 위한 방안에 대하여 논의한다. 네트워크 거리는 삼각형 부등식 성질(triangular inequality property)을 만족하지 못하므로 기존의 기법들은 네트워크 거리를 기반으로 하는 인덱스를 사용하지 않았다. 이러한 기법들은 질의 처리 시 심각한 성능 저하의 문제를 가진다. 사전 계산된 네트워크 거리를 이용하는 또 다른 기법은 저장 공간의 오버헤드가 크다는 문제를 갖는다. 본 논문에서는 이러한 두 가지 문제점들을 동시에 해결하기 위하여 객체들 간의 네트워크 거리를 근사하여 객체들에 대한 인덱스를 구축하고, 이를 이용하여 k-최근접 이웃 질의를 처리하는 새로운 기법을 제안한다. 실제 도로 네트워크를 이용한 정확도 검증 실험을 통하여 제안된 기법의 우수성을 규명한다.

### 1. 서론

도로 네트워크 데이터베이스에서 사용되는 질의들은 k-최근접 이웃 질의(k-nearest neighbor queries), 영역 질의(range queries), 공간 조인 질의(spatial join queries) 등이 있다[1]. 본 논문에서는 도로 네트워크 데이터베이스에서 질의 점에서 네트워크 거리가 가장 가까운 k개의 정적 객체를 검색하는 k-최근접 이웃 질의에 관하여 논의한다. 여기서, 두 객체간의 네트워크 거리란 도로 네트워크에서 임의의 두 객체간의 최단 경로 상에 존재하는 도로 세그먼트 길이들의 총 합이다[1].

기존 유클리드 공간상에서는 임의의 두 객체의 절대 위치만으로도 그 객체들 간의 유클리드 거리를 계산할 수 있다. 그러나 임의의 두 점간의 네트워크 거리는 두 점의 절대 위치만을 가지고 계산할 수 없다. 그 이유는 절대 위치가 같아도 두 점간의 네트워크의 형태에 따라 거리가 달라지기 때문이다. 이러한 이유로 현재까지 정확한 네트워크 거리를 효율적으로 구하기 위한 다양한 방법들이 연구되어 왔으며, 대표적인 예로는 IER 기법, INE 기법, VN3 기법 등이 있다[1][2]. 이러한 기존 기법들은 질의 처리를 위하여 네트워크 거리를 기반으로 하는 객체 인덱싱을 시도하지 않았다. 그 이유는 네트워크 거리가 삼각 부등식(triangular inequality)을 만족시키지 못하는 거리 함수이기 때문이다. 삼각 부등식을 만족시키지 못하는 거리 함수를 사용하는 인덱싱 방법은 질의 처리 시 착오 기각(false dismissal)을 일으킬 수 있다[3]. 그러나 만약 인덱스를 사용할 수 있다면, 비교적 작은 공간 오버헤드만으로도 빠르게 질의를 처리하는 것이 가능하게 된다. 본 논문

에서는 네트워크 거리를 기반으로 정적 객체를 근사 인덱싱(approximate indexing)하고, 이를 이용하여 효과적으로 질의를 처리하는 새로운 기법을 제안한다.

### 2. 제안하는 기법

전체적인 근사 인덱싱 기법에 대해 설명하면 다음과 같다. 우선, 모든 정적 객체를 m차원 유클리드 공간상의 절대 위치로 매핑한다. 이 때, 절대 위치 설정 기준은 m차원 유클리드 공간상의 임의의 두 정적 객체간의 유클리드 거리가 원래의 네트워크 공간상에서의 실제 네트워크 거리를 가능한 유지하도록 하는 것이다. 이후 매핑된 정적 객체의 위치를 다차원 공간 인덱스인 R\*-트리를 이용하여 인덱싱한다. 질의 점 역시 다른 모든 정적 객체와의 네트워크 거리를 가급적 그대로 유지하도록 m차원 유클리드 공간상의 한 점으로 매핑한다. 정적 객체들과 질의 점을 m차원 유클리드 공간상의 절대 좌표로 매핑하기 위한 다양한 방법들이 존재한다[4]. 본 연구에서는 매핑 시 발생하게 되는 오차가 적으면서, 빠르게 매핑할 수 있는 FastMap 알고리즘을 사용한다[5].

매핑하기 위해서는 우선 모든 정적 객체 쌍 간의 네트워크 거리 정보가 필요하다. FastMap은 이 네트워크 거리 정보를 이용하여 각 정적 객체를 m차원 유클리드 공간상의 절대 좌표 점으로 매핑한다. 정적 객체 쌍 간의 네트워크 거리는 다익스트라 알고리즘(Dijkstra algorithm)을 이용하여 쉽게 구할 수 있다. 그러나 네트워크 거리는 FastMap이 요구하는 거리 함수의 세 가지 조건들[5] 중 대칭성 조건과 삼각 부등식 조건을 만족하지 못한다. 따라

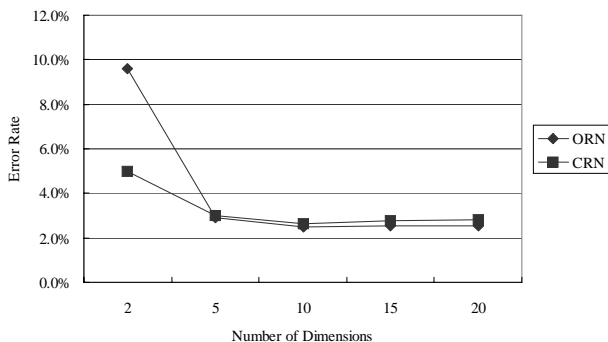
서 네트워크 거리를 FastMap에 바로 적용하는 것은 불가능하다. 본 논문에서는 도로 네트워크에서 두 정적 객체간의 왕복의 거리 차가 크지 않다는 점에 착안하여 두 정적 객체 간 왕복 네트워크 거리의 평균을 두 정적 객체의 네트워크 거리 대신 사용함으로써 이 문제를 해결하였다.

도로 네트워크 위의 임의의 위치에 질의 점이 주어졌을 때, 제안한 인덱스를 이용하여 질의를 처리하기 위해서는 정적 객체가 매핑된  $m$ 차원 유클리드 공간상에서의 질의 점의 절대 좌표를 구해야 한다. 이를 위해서는 질의 점에서 각 피벗 객체(pivot object)[5]까지의 네트워크 거리 정보가 필요하다. 그러나 질의 처리 시점에 질의 점에서  $2 * m$ 개의 피벗 객체까지의 평균 네트워크 거리를 계산하는 것은 매우 큰 처리 비용을 요구한다. 본 논문에서는 도로 네트워크상의 각 노드에서 각 피벗 객체까지의 거리를 미리 계산하여 유지함으로써 이 문제를 해결하였다.

근사 인덱싱 기법의 문제점은 착오기각이 발생할 수 있다는 것이다. 이 문제를 해결하기 위한 방안으로  $k'(\geq k)$ 개의 정적 객체를 검색함으로써 착오기각을 완화시킬 수 있다. 그 이유는 검색할 정적 객체의 수를  $k'$ 개로 늘려 줌에 따라  $k$ 개의 정적 객체를 검색하는 경우에는 착오기각 되었던 정답이 검색 결과에 포함될 확률이 높아지기 때문이다. 따라서 착오기각의 문제점을 크게 완화시킬 수 있다.

### 3. 정확도 검증 실험

본 논문에서는 실험을 통하여 제안된 기법에서 절대 좌표 매핑 시 발생하는 오차를 측정하기 위하여 질의 점으로부터 임의의 두 객체까지 거리의 순서가 바뀌는 정도를 측정하였다. 실험을 위한 실제 도로 네트워크 데이터로서 Oldenburg Road Network(ORN)과 California Road Network(CRN)을 사용하였다. ORN은 노드의 수가 6,105개이고, 도로 세그먼트의 수가 7,034개인 도로 네트워크이다. CRN은 노드의 수가 21,047개이고, 도로 세그먼트의 수가 21,691개인 도로 네트워크이다. 도로 네트워크 내의 정적 객체의 수는 ORN과 CRN에서 각각 도로 세그먼트 수의 0.5%를 취하였으며, 이들의 위치는 전체 도로 네트워크 내에서 균등하게 분포하도록 하였다.



(그림 1) 변환 차원에 따른 오차율.

그림 1은 그 실험 결과를 나타낸 것이다. 각 데이터 집합의 오차율은 ORN에서 최소 2.5%이었으며, CRN에서 최소 2.6%이었다. 차원이 증가함에 따라 대체로 오차율은 감소하였다. 이는 차원의 증가로 인하여 모든 정적 객체간의  $m$ 차원 유클리드 공간상의 상대적 거리가 도로 네트워크 공간상의 상대적 거리를 잘 반영하였기 때문이다. 그러나 특정 차원 수 이후로는 오차율이 더 이상 줄어들지 않는 경향을 보였다. 이는 차원 수가 높아지더라도 FastMap이 가지고 있는 고유의 오차가 크게 개선되지 않기 때문이다. 이 결과를 통하여 적은 수의 차원만을 이용하여 도로 네트워크 공간상의 임의의 객체들을 표현 가능하다는 것을 알 수 있다.

### 4. 결론

본 논문에서는 도로 네트워크 데이터베이스에서 정적 객체들 간의 네트워크 거리를 기반으로 하는 근사 인덱싱 및 질의 처리 기법을 제안하였다. 제안하는 기법은  $k$ -최근접 이웃 질의뿐만 아니라 영역 질의와 공간 조인 질의 등 다양한 질의 처리에 활용될 수 있다. 정확도 검증 실험을 통하여 제안하는 근사 인덱싱 기법의 정확성을 검증하였다. 제안된 기법은 도로 네트워크 데이터베이스에서 인덱스를 기반으로 하는 질의 처리를 가능하게 함으로써 수용 가능한 저장 공간의 오버헤드만을 가지고 질의 처리 성능을 크게 개선시킬 수 있는 기반을 마련하였다는 점에서 큰 의미가 있다.

### 감사의 글

본 논문은 제주대학교를 통한 정보통신부 및 정보통신진흥원의 대학 IT연구센터 지원사업의 지원을 받았습니다. (IITA-2005-C1090-0502-0009)

### 참고문헌

- [1] D. Papadias et al., "Query Processing in Spatial Network Databases," In *Proc. Int'l. Conf. on Very Large Data Bases, VLDB*, pp. 802-813, 2003.
- [2] M. Kolahdouzan and C. Shahabi, "Voronoi-Based K-Nearest Neighbor Search for Spatial Network Databases," In *Proc. Int'l. Conf. on Very Large Data Bases, VLDB*, pp. 840-851, 2004.
- [3] B. Yi, et al, "Efficient Retrieval of Similar Time Sequences Under Time Warping," In *Proc. IEEE Int'l. Conf. on Data Engineering, ICDE*, pp. 201-208, 1998.
- [4] J. Wang et al., "Evaluating a Class of Distance-Mapping Algorithms for Data Mining and Clustering," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data Mining, ACM SIGKDD*, pp. 307-311, 1999.
- [5] C. Faloutsos and K. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," In *Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD*, pp. 163-174, 1995.