Query Processing in Spatial Network Databases

Instructor: Cyrus Shahabi
Outline

• Introduction
• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Introduction

- Euclidean distance

Where is the nearest Starbucks?????
Introduction

- Euclidean distance

Where is the nearest Starbucks ??????
Introduction

- Euclidean distance

Only 8 miles... yaaaayyy!!

Closest...
Introduction

- Network distance

Considering the underlying road network, is the nearest Starbucks still the same?
Introduction

- Network distance

Closest...

NO !!!
Introduction

- Euclidean distance vs. Network distance

Any relationship?

\[ d_E \leq d_N \]

the Euclidean distance between two points is equal or smaller than their network distance.
Introduction

- How can we represent a road network?
- Graphs
Outline

• Introduction
• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Disk-based graph representations

- A graph can be represented as
  - Two-dimensional matrix
  - An adjacency list
Disk-based graph representations: 2D Matrix

- Two-dimensional matrix

- Disadvantage?

Adjacency matrix

<table>
<thead>
<tr>
<th></th>
<th>n1</th>
<th>n2</th>
<th>n3</th>
<th>n4</th>
<th>n5</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>n2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>n3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>n4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>n5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Disk-based graph representations: 2D Matrix

– Two-dimensional matrix

– Disadvantage?
  Sparse, More I/O

Adjacency matrix

<table>
<thead>
<tr>
<th></th>
<th>n1</th>
<th>n2</th>
<th>n3</th>
<th>n4</th>
<th>n5</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>n2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>n3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>n4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>n5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Disk-based graph representations: CCAM structure

• The Connectivity-clustered Access Method (CCAM) structure
  – Each node has a list that stores its neighbors
  – Stores the lists of neighbor nodes together
Disk-based graph representations: CCAM structure

• An example
Disk-based graph representations: CCAM structure

- An example

A graph

Index: a B-tree in order of node id

Disk: Adjacency lists

Adjacency list of n1

null
Disk-based graph representations- CCAM structure

• CCAM vs. 2D Matrix?
• CCAM is preferable for applications, such as road networks, where the graphs are sparse.
Outline

• Introduction
• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Spatial query processing in Euclidean Space

- An R-tree index
  - Multidimensional extension of B-tree
Outline

• Introduction
• Related work
  – Disk-based graph representations- 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Architecture

- Underlying network and spatial entities are separated
- Index the entity datasets (e.g., hotels) separately by R-trees (called Objects R-tree)
- For the network: preserve connectivity (adjacency comp) & location (network R-tree)
Functions

- **check_entity(seg, p)**: is a Boolean function that returns true if point (entity) \( p \) lies on the network segment \( seg \) (i.e., \( seg \) covers \( p \))
  - MBR of \( seg \) is used for filtering and its poly-line representation for refinement.

- **find_segment(p)**: outputs the segment that covers point \( p \) by performing a point location query on the network R-tree.

- **find_entities(seg)**: returns entities covered by segment \( seg \).
  - first finds all the candidate entities that lie in the MBR of \( seg \), and then eliminates the false hits using the poly-line of \( seg \).

- **compute_ND(p1, p2)**: returns the network distance of two arbitrary points \( p1, p2 \) in the network,
  - by applying a (secondary-memory) Dijkstra's algorithm to compute the shortest path from \( p1 \) to \( p2 \).
Nearest Neighbor - IER

- Example: NN of q?

- Step 1: Find Euclidean NN $p_{E1}$ of q on the entity R-tree
- Step 2: Compute the network distance $d_N(q, p_{E1})$ of $p_{E1}$
- Step 3: Euclidean lower-bound. Objects closer to q than $p_{E1}$ should be within Euclidean distance $d_{E_{max}} = d_N(q, p_{E1})$ from q. Only check SHADED AREA!
Nearest Neighbor - IER

- Example: NN of q?

- Step 1: Find Euclidean NN $p_{E1}$ of q on the entity R-tree
- Step 2: Compute the network distance $d_{N}(q, p_{E1})$ of $p_{E1}$
- Step 3: Euclidean lower-bound. Objects closer to q than $p_{E1}$ should be within Euclidean distance $d_{E_{\text{max}}} = d_{N}(q, p_{E1})$ from q. Only check SHADED AREA!

- Step 4: Find 2nd Euclidean NN $p_{E2}$ within range $d_{E_{\text{max}}}$.
- Step 5: $d_{N}(q, p_{E2}) < d_{N}(q, p_{E1})$. Current NN is $p_{E2}$.
- Step 6: Set $d_{E_{\text{max}}} = d_{N}(q, p_{E2})$. Range is now smaller.
- Step 7: Next NN $p_{E3}$ is out of $d_{E_{\text{max}}}$ range. Terminate!
IER: Shortcoming

- IER performs well if the ranking of the data points by their Euclidean distance is similar to that with respect to the network distance. Otherwise, a large number of Euclidean NNs may be inspected before the network NN is found.

- E.g., P5 is the closest in network distance but farther than p1 to p4 in Euclidean distance.

Since $p5$ has the largest Euclidean distance, it will be examined after all other entities, i.e., $p1$ to $p4$ correspond to false hits, for which the network distance computations are redundant.
Nearest Neighbor - IER

- Example with Numbers: NN of q?

  - Step 1: 1st Euclidean NN of q is $p_{E1}$. $d_E(q, p_{E1}) = 50$
  - Step 2: Compute the network distance $d_N(q, p_{E1}) = 100$. Hence $d_{E_{max}} = 100$.
  - Step 3: Objects closer to q than $p_{E1}$ should be within a Euclidean range of 100.
Nearest Neighbor - IER

• Example with Numbers: NN of q ?

• Step 1: 1st Euclidean NN of q is $p_{E1}$. 
  $d_E(q, p_{E1}) = 50$

• Step 2: Compute the network distance 
  $d_N(q, p_{E1}) = 100$. Hence $d_{E_{max}} = 100$.

• Step 3: Objects closer to q than $p_{E1}$ should be within a Euclidean range of 100.

• Step4: 2nd Euclidean NN of q is $p_{E2}$. 
  $d_E(q, p_{E2}) = 75$

• Step 5: Compute the network distance 
  $d_N(q, p_{E2}) = 90$. $d_N(q, p_{E2}) < d_N(q, p_{E1})$. Current NN is $p_{E2}$.

• Step 6: Set $d_{E_{max}} = 90$. Range is now smaller.

• Step 7: Next NN $p_{E3}$ is out of $d_{E_{max}}$ range. Terminate!
Nearest Neighbor – INE

- Incremental Network Expansion (INE) performs network expansion (starting from $q$), and examines entities in the order they are encountered.

- $n_1$, $n_2$, $n_3$, $n_4$, $n_5$
- $q$
- $p_1$, $p_2$

- -> Network node
- -> Object (entity)
- -> Query point
Nearest Neighbor – INE

Find NN for a given query point $q$. 

Min Heap
Nearest Neighbor – INE

Find the edge that q is on using Network R-tree, initiate heap

Min Heap
(n₁, 2), (n₅, 3)
Nearest Neighbor – INE

For each edge $e$ in heap, find entities that are on $e$ and add $e$’s neighboring edges into heap.
Nearest Neighbor – INE

Add $n_2$ and $n_4$ into heap

Min Heap

$(n_5, 3), (n_2, 4), (n_4, 9)$
Nearest Neighbor – INE

Add $n_4$ into heap

Min Heap

$(n_2, 4), (n_4, 7), (n_4, 9)$
Nearest Neighbor – INE

- $d_N(q, p_1) = d_{N_{\text{max}}} = 4$
- The next entry in the heap $n_2$ doesn’t have a smaller distance than $d_{N_{\text{max}}}$. Thus, the algorithm terminates.

$\begin{align*}
\text{Min Heap} \\
(n_2, 4), (n_4, 7)
\end{align*}$

$p_1$ is found on the edge $(n_5, n_4)$
Outline

• Introduction
• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Range Queries – ER

• Range query: given a query point $q$, a range $e$ and a spatial dataset $S$, find all objects that are within network distance $e$ from $q$.

• **Range Euclidean Restriction (RER) method:**
  
  – Perform a range query at the entity dataset and find the set of objects $S'$ within (Euclidean) distance $e$ from $q$.
  
  • $S'$ is guaranteed to avoid false misses using lower-bound property ($d_N(q,p) \leq e \Rightarrow d_E(q,p) \leq e$).
  
  • But it may contain a large number of false hits.
  
  – RER examines all segments within network distance $e$ from $q$. Points of $S'$ that fall on some segment, are removed from $S'$ and returned to the user as result.
  
  – The process terminates when all the segments in the range are exhausted, or when $S'$ becomes empty.
Range Queries – NE

- **Range Network Expansion (RNE) algorithm**
  - First compute the set $QS$ of qualifying segments within network range $e$ from $q$.
  - Retrieve the data entities falling on these segments (intersection-join between $QS$ and objects)

Problem? $QS$ may be large
Start traversing the object R-tree from root.

1. Compute $QS_i$ for each entry $E_i$ in the current R-tree node.
   - Ex: $QS_1 = \{\}$, $QS_2 = \{\text{all segments except } n_5 n_8 \text{ and } n_1 n_4\}$, $QS_5 = \{n_q n_2, n_2 n_5, n_2 n_6\}$
   - For each entry $E_i$
     - If $QS_i \neq \{\}$, recurse down the tree
2. If current node is a leaf (suppose $E_6$), its points only be checked against $QS_6$.
3. Return entities falling on the segments in $QS_i$ ($QS_6$). Note that $c$ is not a qualifying object.
Outline

• Introduction
• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree
• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method
• Experiments
• Summary
Experiments - NN queries

- IER (Incremental Euclidean Restriction) vs. INE (Incremental Network Expansion).
- Cost as a function of the ratio entity/edge cardinality
- Number of neighbours to be retrieved $k=10$
Experiments - Range Search

• RER (Range Euclidean Restriction) vs. RNE (Range Network Expansion).
• Cost as a function of the ratio entity/edge cardinality
• Length of the range $e=1\%$ of the data universe side length

\[\text{Page Accesses}\]

\[\text{CPU time - msecs}\]

\[\text{cardinality ratio - } |S|/|N|\]
Outline

• Introduction

• Related work
  – Disk-based graph representations: 2DMatrix, CCAM structure
  – R-tree

• Spatial query in network databases
  – Architecture
  – Spatial queries:
    • Nearest neighbor query
      – Incremental Euclidean Restriction (IER) method
      – Incremental Network Expansion (INE) method
    • Range query
      – Range Euclidean Restriction (RER) method
      – Range Network Expansion (RNE) method

• Experiments

• Summary
Summary

• Network distance is a more realistic metric than Euclidean distance.

• *Euclidean restriction* assumes the lower bounding property, which may not always hold in practice (if, for instance, the edge cost is defined as the expected travel travel time). On the contrary, *network expansion* permits a wide variety of costs associated with the edges.

• *Network expansion* has superior performance for nearest neighbour and range queries.
References

• Dimitris Papadias, Jun Zhang, Nikos Mamoulis, Yufei Tao: Query Processing in Spatial Network Databases. VLDB 2003:802-813

• A presentation by Afsin Akdogan in csci587 Fall’2010