Voronoi-Based K Nearest Neighbor Search for Spatial Network Databases

Instructor: Cyrus Shahabi
Agenda

- Problem Definition
- Related Work
- Voronoi Diagrams
- Voronoi Based Network Nearest Neighbor (VN^3)
- Experiments
- Discussion
Problem Definition

**kNN Problem:**
Given a set of objects $P$ and query point $q$, find the $k$ objects in $P$ that are closest to $q$.
Related Work

- NN Query: Roussopoulos et al., SIMGOD 1995
- K-NN for moving query point: Zong et al., SSTD 2001
- Time-parameterized queries: Tao et al., SIMGOD 2002
- Continuous NN Search: Tao et al., VLDB 2002

Object Based
Related Work

- Sina-Continuous querying: Mokbel et al., SIGMOD 2004
- Sea-CNN queries: Xiong et al., ICDE 2005
- Monitoring kNN on moving objects: Yu et al., ICDE 2005
- Conceptual partitioning: Mouratidis et al., SIGMOD 2005

Space Based
Example 1:
Finding the 3 closest shopping centers
Example 2:
Finding the 3 closest restaurants to USC with Yahoo
Related Work

- Road Network Embedding for K-NN Search: Shahabi et al., ACM-GIS 2002
- Query processing in SNDB: Papadias et al., VLDB 2003
- Voronoi-based kNN in SNDB: Kolahdouzan et al., VLDB 2004
Related Work

Query processing in SNDB: Papadias et al., VLDB 2003

Incremental Network Expansion (INE)
- Blind expansion hence redundant node access
Related Work

Query processing in SNDB: Papadias et al., VLDB 2003

– Incremental Network Expansion
Preliminaries: Voronoi Diagram

- Given set of sites (POI), a Voronoi diagram partitions the plane into disjoint Voronoi polygons.
- The region including a site \( p \) includes all locations which are closer to \( p \) than to any other object \( p' \).

\[ q \text{ is closer to the generator of the Voronoi Polygon than any other generator} \]
Preliminaries: Voronoi Properties

• Property 1: Voronoi diagram is unique
• Property 2: Voronoi edges are
  – shared by two generators
  – in equal distance to neighboring generators
• Property 3: Nearest generator of p is among the generators whose VP shares similar edges with p
• Property 4: Average number of Voronoi edges per VP is at most 6
Network Voronoi Diagram

Border Point: equal network distance to adjacent generators

Network Voronoi Polygon

Network Edge
VN$^3$ Approach

- **Offline Index Generation**
  1. Network Voronoi Construction
  2. Index Generation (R-tree)
  3. Distance Precomputation

- **Online Query Processing**
  1. Find 1$^{st}$ NN
  2. Find k NN-> Filter & Refine
Overview

• How to compute distance from q to candidates?
• Break distance from q to any $P_i$ to:
  1. Distance from q to its borders
     I. Online Progressive Expansion
     II. Offline Pre-calculation
  2. Distance from borders of q to borders of $P_i$
     I. NVP Expansion
     II. Distance Computing Optimization
Offline Step

- Compute Network Voronoi Polygons (NVP)
- Index NVPs with R-Tree
- Precompute the intra (across polygons) and inter (within polygons) network distances for each NVP

1. Precompute the Shortest Path distance from each generator to its border point
2. Precompute the Shortest Path between border points

Example: $d_n(q, P_9) = \min(d_n(q, b_{34}) + d_n(b_{34}, P_9), d_n(q, b_{35}) + d_n(b_{35}, P_9), d_n(q, b_{36}) + d_n(b_{36}, P_9))$

$P_9$

\[ b_{34} \quad b_{35} \quad b_{36} \quad \ldots \quad b_1 \quad b_2 \quad b_{15} \quad b_{13} \quad \ldots \]

\[ P_9 \quad d_n(b_{34}, P_9) \quad d_n(b_{35}, P_9) \quad \ldots \quad d_n(b_{1}, b_2) \quad \ldots \quad d_n(b_{1}, b_{13}) \quad \ldots \]
Online Step

- Find 1
dst NN
  - Search R-Tree to find the NVP that overlaps q
  - Report the generator of the NVP as the 1
  st NN;

Cost: $O(\log n)$
Online Step

- 1\textsuperscript{st} NN = P_1
- 2\textsuperscript{nd} NN ∈ \{ Neighbors( P_1 ) \}
  (e.g., P_3)
- 3\textsuperscript{rd} NN ∈ \{ Neighbors( P_1 ) \cup Neighbors( P_3 ) \}
  (e.g., P_2)
- 4\textsuperscript{th} NN ∈ \{ Neighbors( P_1 ) \cup Neighbors( P_3 ) \cup Neighbors( P_2 ) \}
- ...

GOOD NEWS! In theory average # of neighbors = O(5k+1)=O(k)
In Practice:  << 5 \times K
Performance of VN$^3$

• Data set:
  – Road network in Los Angeles (from NavTeq)
    • 110,000 streets, 79,800 intersections
    • Different points of interest:
      – restaurants, auto services, schools, parks, shopping centers, hospitals

• Measured:
  – Query response time and CPU
    • Comparison with INE [Papadias et al. (vldb03)]
  – Size of candidate set of VN$^3$ filter
    • Comparison with R-tree-based KNN [Seidl et al. (sigmod98) and Hjaltason et al. (tods99)]
### Pre-computation Overhead of VN³

<table>
<thead>
<tr>
<th>Entities</th>
<th>Cardinality</th>
<th>Points Inside each NVP</th>
<th>Average BPs per NVP</th>
<th>No. of Pre-comp. Border to Border</th>
<th>No. of Pre-comp. Border to inside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitals</td>
<td>0.04%</td>
<td>1698</td>
<td>52</td>
<td>232,000</td>
<td>8,781,000</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.16%</td>
<td>458</td>
<td>25</td>
<td>225,600</td>
<td>4,653,000</td>
</tr>
<tr>
<td>Parks</td>
<td>0.53%</td>
<td>142</td>
<td>14</td>
<td>239,500</td>
<td>2,630,000</td>
</tr>
<tr>
<td>Schools</td>
<td>1.15%</td>
<td>64</td>
<td>10</td>
<td>246,000</td>
<td>1,787,000</td>
</tr>
<tr>
<td>Auto. Svc.</td>
<td>3.26%</td>
<td>38</td>
<td>7</td>
<td>239,900</td>
<td>1,611,000</td>
</tr>
<tr>
<td>Restaurants</td>
<td>5.80%</td>
<td>27</td>
<td>6</td>
<td>243,600</td>
<td>1,348,000</td>
</tr>
</tbody>
</table>

- Smaller/more NVPs when dense
- Larger/less NVPs when sparse
- Naive approach: 3.2 billion pre-computations
  - Increases when sparse
  - Almost same

# neue approach
Comparison of VN³ and INE

- $K = 1$

<table>
<thead>
<tr>
<th>Entities</th>
<th>Density, Cardinality</th>
<th>VN³ (cpu), db sec</th>
<th>INE (cpu), disk sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitals</td>
<td>0.04%, 46</td>
<td>(0), 0.018</td>
<td>(0.30), 12.4</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.16%, 173</td>
<td>(0), 0.020</td>
<td>(0.090), 3.6</td>
</tr>
<tr>
<td>Parks</td>
<td>0.53%, 561</td>
<td>(0), 0.021</td>
<td>(0.030), 1.4</td>
</tr>
<tr>
<td>Schools</td>
<td>1.15%, 1230</td>
<td>(0), 0.027</td>
<td>(0.015), 0.6</td>
</tr>
<tr>
<td>Auto. Svc.</td>
<td>3.26%, 2093</td>
<td>(0), 0.030</td>
<td>(0.013), 0.67</td>
</tr>
<tr>
<td>Restaurants</td>
<td>5.80%, 2944</td>
<td>(0), 0.032</td>
<td>(0.010), 0.49</td>
</tr>
</tbody>
</table>

A substantial increase for sparse
Comparison of VN$^3$ and INE

- **K = 5**

<table>
<thead>
<tr>
<th>Entities</th>
<th>Density, Cardinality</th>
<th>VN$^3$ (cpu), db sec</th>
<th>INE (cpu), disk sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitals</td>
<td>0.04%, 46</td>
<td>(1.5), 6.5</td>
<td>(1.7), 78.3</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.16%, 173</td>
<td>(0.45), 3.3</td>
<td>(0.5), 21.1</td>
</tr>
<tr>
<td>Parks</td>
<td>0.53%, 561</td>
<td>(0.15), 1.5</td>
<td>(0.2), 8.2</td>
</tr>
<tr>
<td>Schools</td>
<td>1.15%, 1230</td>
<td>(0.06), 0.75</td>
<td>(0.07), 3.5</td>
</tr>
<tr>
<td>Auto. Svc.</td>
<td>3.26%, 2093</td>
<td>(0.01), 0.47</td>
<td>(0.05), 2.43</td>
</tr>
<tr>
<td>Restaurants</td>
<td>5.80%, 2944</td>
<td>(0.01), 0.87</td>
<td>(0.03), 1.5</td>
</tr>
</tbody>
</table>

- Excluding K=1, VN$^3$ is between 2 (high densities / larger K) to 14 (low densities / larger K) times faster than INE.
## VN³ vs INE

VN³ finds the 1\textsuperscript{st} NN using R-Tree
INE needs to expand the network around q

### CPU Time:
INE uses a queue which is incrementally updated

### Table: Query Processing Time (Sec)

<table>
<thead>
<tr>
<th>Entities</th>
<th>Qty (density)</th>
<th>K=1</th>
<th>K=5</th>
<th>K=10</th>
<th>K=25</th>
<th>K=50</th>
<th>K=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital</td>
<td>46 (0.0004)</td>
<td>Vn³</td>
<td>VN³</td>
<td>VN³</td>
<td>VN³</td>
<td>VN³</td>
<td>VN³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.018</td>
<td>12.4</td>
<td>6.5</td>
<td>78.3</td>
<td>14.0</td>
<td>165.1</td>
</tr>
<tr>
<td>Shopping Centers</td>
<td>173 (0.0016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.020</td>
<td>3.6</td>
<td>3.3</td>
<td>21.1</td>
<td>6.9</td>
<td>44.0</td>
</tr>
<tr>
<td>Parks</td>
<td>561 (0.0053)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.021</td>
<td>1.4</td>
<td>1.5</td>
<td>8.2</td>
<td>6.9</td>
<td>44.0</td>
</tr>
<tr>
<td>Schools</td>
<td>1230 (0.0115)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.027</td>
<td>0.6</td>
<td>0.75</td>
<td>3.5</td>
<td>6.9</td>
<td>44.0</td>
</tr>
<tr>
<td>Auto Services</td>
<td>2093 (0.0326)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.030</td>
<td>0.57</td>
<td>0.65</td>
<td>2.43</td>
<td>6.9</td>
<td>44.0</td>
</tr>
<tr>
<td>Restaurants</td>
<td>2944 (0.0580)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
<td>Vn³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
<td>(cpu)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
<td>disk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.032</td>
<td>0.49</td>
<td>0.57</td>
<td>1.34</td>
<td>6.9</td>
<td>44.0</td>
</tr>
</tbody>
</table>
vn³: performance of filter step

- usgs data, euclidean space

- ratio of size of candidate set to k decreases as k increases
  - some of the new neighbors are already explored!

- vn³'s filter behaves independently from density/distribution of points of interest
  - size of the candidate set (o(k)) is independent of the density/distribution

- better selectivity as compared to other approaches (up to 4 times)
- very low variance (predictable behavior)
Conclusion

• A novel approach for KNN queries in SNDB
• Based on:
  – Pre-calculating the solution space (first order Voronoi diagrams)
  – Pre-computing some distances (from borders to points inside each polygon)
• Outperforms the only other solution for SNDB
• Independent from object distribution
• Outperforms the solutions for Euclidean space in “filtering”
Problem Definition

• Given a NVD and query location $q$, how can we efficiently find the NVP that overlaps with $q$; $\text{contain}(q)$?
Indexing NVD

• Voronoi R-tree (VR-tree)
  – Generate NVPs by connecting border points (See Section 4.3)
  – Create minimum bounding rectangle (MBR) around NVPs and index MBRs with R-Tree

VR-tree [Shahabi’04, Taniar’07, Safar’06, Xuan’10]
Indexing NVD

• VR-tree – Problem 1
  – VR-tree splits the network space with hierarchically nested and largely over-lapping MBRs
  – Non-disjoint partitioning of the space
Indexing NVD

- **VR-tree – Problem 2**
  - Although the NVPs are computed based on the network distance metric, they are treated as polygons in Euclidean space.
  - Misclassification of edges: False-negative edges.

![Diagram illustrating VR-tree and NVP with marked false-negative edges.](image-url)
Indexing NVD

• VQ-tree
  – NVPs are
    • mutually exclusive: they do not have any overlapping areas
    • collectively exhaustive: every node/edges in the network is associated with at least one generator
  – Disjoint decomposition of the NVPs
Indexing NVD

- VQ-tree
  - NVPs are
    - mutually exclusive: they do not have any overlapping areas
    - collectively exhaustive: every node/edges in the network is associated with at least one generator
  - Disjoint decomposition of the NVPs
Indexing NVD

• VQ-tree
  – NVPs are
    • mutually exclusive: they do not have any overlapping areas
    • collectively exhaustive: every node/edges in the network is associated with at least one generator
  – Disjoint decomposition of the NVPs
Indexing NVD

• VQ-tree
  – Each network NVP consists of disjoint quad-tree blocks
Indexing NVD

- **VQ-tree**
  - Each network NVP consists of disjoint quad-tree blocks
  - Leaf nodes do not keep any information about the network but store the region information (i.e., coordinates) of the quad-blocks
Discussion

• What if the edge weights are changing?
• What if the both query and data objects are moving?
• What if you like to find the nearest hotel and gas station at the same time?
References

- A presentation by Ugur Demiryurek in csci587 Fall’2010