

Reverse kNN search in Arbitrary Dimensionality

Seyed Jalal Kazemitabar

Original paper by Y. Tao, D. Papadias, and X. Lian

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Nearest Neighbor Queries

What are the two nearest stars to Andromeda?



Where is the nearest restaurant?



Where is the nearest....



InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Algorithms for finding NN

- Elementary methods:
 - Search Algorithm + Indexing Data Structure = NN solution
 - Search Algorithm: BF, DFS
 - Indexing Data Structure: R-tree, R*-tree
- More advanced methods:
 - Search Algorithm + Branch & Bound Methods + Indexing Data Structure = NN solution
 - Branch & Bound Methods: Mindist, Maxdist, Minmaxdist

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Reverse Nearest Neighbors Queries

What are the fireplaces I'm nearest to?



Which houses I'm the closest restaurant to?

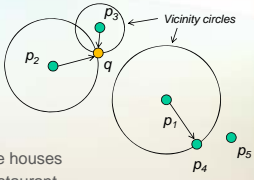


InfoLab.usc.edu Geospatial Information Management (Fall 2009)

RNN Definition

- A data point p is the reverse nearest neighbor of query point q , if there is no point p' such that $dist(p', q) < dist(p, q)$, i.e. q is the NN of p .

$NN(p_2) = NN(p_3) = q$
 $RNN(q) = \{p_2, p_3\}$



- In our example, p_2, p_3 are the houses for which q is the nearest restaurant
- Is RNN a symmetric relation?

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Related Work

RNN Algorithms

Main idea: Pre-computing, Filter/refinement

Methods: KM, YL, SAA, SFT

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

KM

- Original RNN method
- For all p :
 1. Pre-compute $NN(p)$
 2. Represent p as a vicinity circle
 3. Index the MBR of all circles by an R-tree (Named RNN-tree)
 4. $RNN(q)$ = all circles that contain q
- Needs two trees: RNN-tree & R-tree

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

YL

- YL:
 - Merges the trees
 - What happens if we insert p_5 ?
 1. $RNN(p_5) = ?$
Find all points that have p_5 as their new NN
 2. Update the vicinity circles of those points in the index
 3. Compute $NN(p_5)$ and insert the corresponding circle in the index
 - Drawbacks?

Techniques that rely on pre-processing cannot deal efficiently with updates

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

SAA

- Elimination of the need for pre-computing all NNs in filter/refinement methods
- SAA:
 - Divide the space around query into six equal regions
 - Find $NN(q)$ in all regions (candidate keys)
 - Either (i) or (ii) holds for each candidate key p
 - (i) p is in $RNN(q)$
 - (ii) No $RNN(q)$ in S_i
 - $RNN(q) = \{p_i\}$
 - Any Drawbacks?

The number of regions increases exponentially with the dimensionality

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

SFT

- Filter
 1. Find the k NNs of the query q (k candidates)
 2. Eliminate the points that are closer to other candidates than q .
 3. Apply *Boolean range queries* to determine the actual RNNs
- A Boolean range query terminates as the first data point is found
- Drawbacks?

False misses
Choosing a proper k

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

- Concluding former methods:

	Dynamic data	Arbitrary dimensionality	Exact result
KM, YL	No	Yes	Yes
SAA	Yes	No	Yes
SFT	Yes	Yes	No

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Half-plane pruning

- Can p' be closer to q than p can be?

- If p_1, p_2, \dots, p_n are n data points, then any node whose MBR falls inside $\bigcup_{i=1..n} P(p_i, q)$ cannot contain any RNN result.

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

□ Pruning an R-tree MBR:

□ Drawbacks?

$O(n^2)$ processing time in terms of bisector trimming for computing N^{res}
 Computation of intersections does not scale with dimensionality

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

□ Approximating the residual MBR

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

□ An MBR can be pruned if its residual region is empty

□ The approximation is a superset of the real residual region

□ We can prune an MBR if its approximate residual is empty

□ Good news:

$O(n)$ processing time for computing N^{resM}
 No more hyper-polyhedrons to make the intersection computation complex

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

TPL Algorithm

□ The big picture

- Uses best-first search
- Utilizes one R-tree as the data structure
- Includes filtering/ refinement phases
- Uses candidate points to prune entries
- Filters visited entries to obtain the set S_{cand} of candidates
- Adds pruned entries to set S_{ref}
- S_{ref} is used in the refinement step to eliminate false hits

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

TPL Example

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Filtering step

Action	Heap	S_{cand}	S_{ref}
Visit root	$\{N_{10}, N_{11}, N_{12}\}$	\emptyset	\emptyset

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
Visit N_{10}	$\{N_9, N_{11}, N_2, N_1, N_{12}\}$	$\{\}$	$\{\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
Visit N_3	$\{N_{11}, N_2, N_1, N_{12}\}$	$\{p_1\}$	$\{p_3\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
Visit N_{11}	$\{N_5, N_2, N_1, N_{12}\}$	$\{p_1\}$	$\{p_3, N_4, N_6\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
Visit N_5	$\{N_2, N_1, N_{12}\}$	$\{p_1, p_2\}$	$\{p_3, N_4, N_6, p_8\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
Visit N_1	$\{N_{12}\}$	$\{p_1, p_2, p_3\}$	$\{p_3, N_4, N_6, p_8, p_9, N_2, p_7\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Action	Heap	Scnd	Srtn
	$\{\}$	$\{p_1, p_2, p_3\}$	$\{p_3, N_4, N_6, p_8, p_9, N_2, p_7, N_{12}\}$

infod.ab.usc.edu Geospatial Information Management (Fall 2009)

Refinement Heuristics

- Let P_{in} be the set of points and N_{in} be the set of nodes in S_{in}
- A point p from S_{out} can be discarded as a false hit if there is a point $p' \in P_{in}$ such that either of the following hold:
 - (i) $dist(p, p') < dist(p, q)$
 - (ii) There is a node MBR $N \in N_{in}$ such that $max_{p' \in P_{in}} dist(p, N) < dist(p, q)$
- A candidate point can be eliminated if it is closer to another candidate point than to the query
- A point p from S_{out} can be reported as an actual result if the following conditions hold:
 - (i) There is no point $p' \in P_{in}$ such that $dist(p, p') < dist(p, q)$
 - (ii) For every node $N \in N_{in}$: $max_{p' \in P_{in}} dist(p, N) > dist(p, q)$
- If none of the above works, visit all node MBRs $N \in N_{in}$ where $min_{p' \in P_{in}} dist(p, N) < dist(p, q)$ and use the mentioned heuristics considering the newly visited entries

InfoLab, USC.edu Geospatial Information Management (Fall 2009)

Action	S_{out}	S_{in}	Actual results
	$\{p_1, p_2, p_3\}$	$\{p_3, N_4, N_5, N_6, N_7, N_{12}\}$	\emptyset
Invalidate p_1	$\{p_2, p_3\}$	$\{N_4, N_5, N_6, N_{12}\}$	\emptyset
Validate p_3	$\{p_3\}$	$\{N_4, N_5, N_6, N_{12}\}$	\emptyset
Remove N_4, N_5	$\{p_3\}$	$\{N_4, N_{12}\}$	$\{p_3\}$

InfoLab, USC.edu Geospatial Information Management (Fall 2009)

Action	S_{out}	S_{in}	Actual results
	$\{p_2\}$	$\{N_4, N_{12}\}$	$\{p_2\}$
Access N_4	$\{p_2\}$	$\{p_4, p_5, N_{12}\}$	$\{p_2\}$
Invalidate p_2	\emptyset	$\{N_{12}\}$	$\{p_2\}$

InfoLab, USC.edu Geospatial Information Management (Fall 2009)

RkNN pruning

- Return all points that have q as one of their k nearest neighbors

- Let $\{p_1, p_2, \dots, p_k\}$ be a subset of $\{p_1, p_2, \dots, p_n\}$. Each of the $\binom{n}{k}$ subsets, prunes the area $\{p_1, p_2, \dots, p_k, q\}$

InfoLab, USC.edu Geospatial Information Management (Fall 2009)

KTPL Algorithm

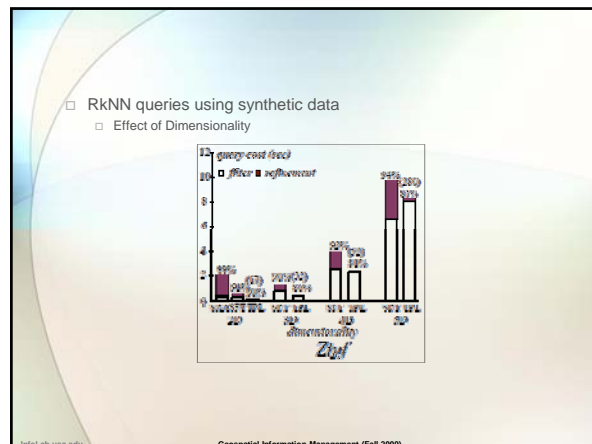
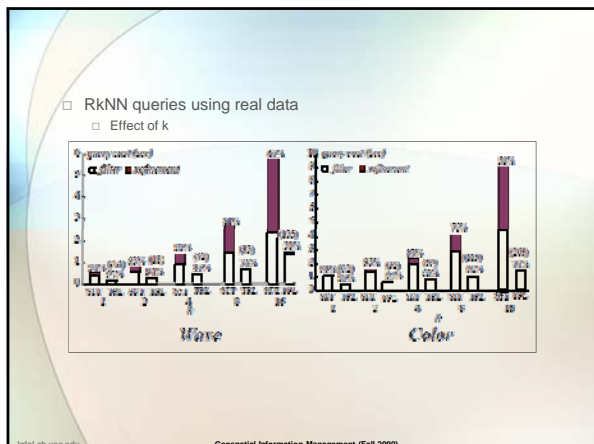
- Same filtering as TPL
- Same refining with the following exceptions:
 - A point can be pruned if k points are found within distance $dist(p, q)$ from p
 - A counter is associated with each point (initialized to k) and decreases when such a point is found
 - A candidate is eliminated if counter = 0
 - No prior knowledge of number of points in a node, so no application of $max_{p' \in P_{in}} dist(p, N) < dist(p, q)$ in pruning
 - A point p can be pruned if a node N is found such that $max_{p' \in P_{in}} dist(p, N) < dist(p, q)$ and $min_{p' \in P_{in}} dist(p, N) \geq counter(p)$

InfoLab, USC.edu Geospatial Information Management (Fall 2009)

Experiments

- RNN queries on real data

InfoLab, USC.edu Geospatial Information Management (Fall 2009)



Conclusion

- TPL is good in that it
 - Supports arbitrary values of k
 - KM, YL, MVZ
 - Can deal efficiently with database updates
 - KM, YL, MVZ
 - Is applicable to data of dimensionality more than two
 - SAA, MVZ
 - Retrieves exact results
 - SFT
 - Results in fast results!

InfoLab.usc.edu Geospatial Information Management (Fall 2009)

References

1. "Reverse kNN Search in Arbitrary Dimensionality". Y. Tao, D. Papadias, X. Lian.
2. <http://202.118.18.45/seminars> a presentation by Guo Peng

InfoLab.usc.edu Geospatial Information Management (Fall 2009)