Nearest neighbor search using the Hierarchical Navigable Small World (HNSW) algorithm

Sebastian Björkqvist

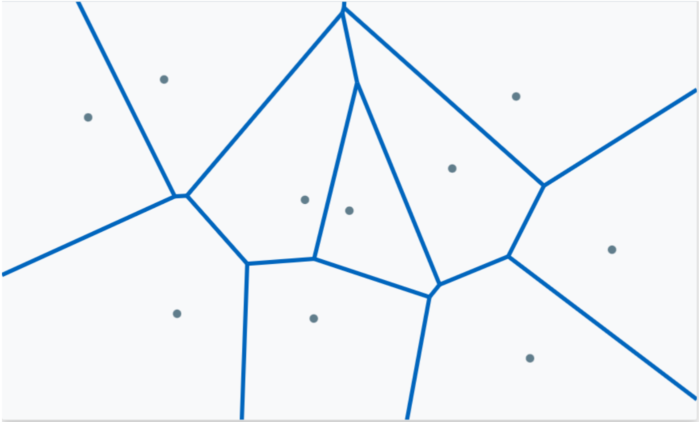
Lead AI Developer, IPRally

May 12, 2023

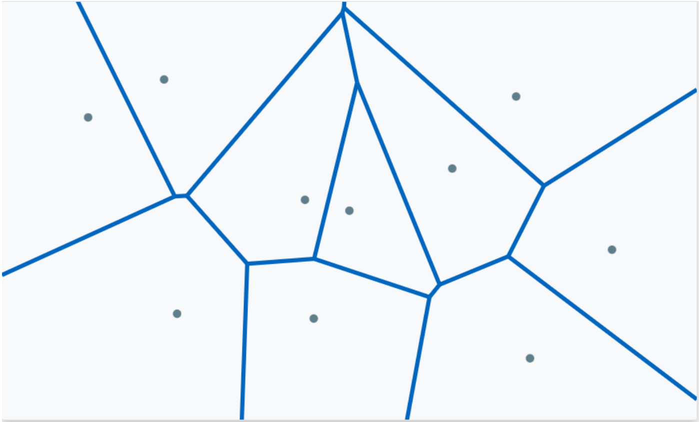
Outline

Voronoi diagram for a set of points

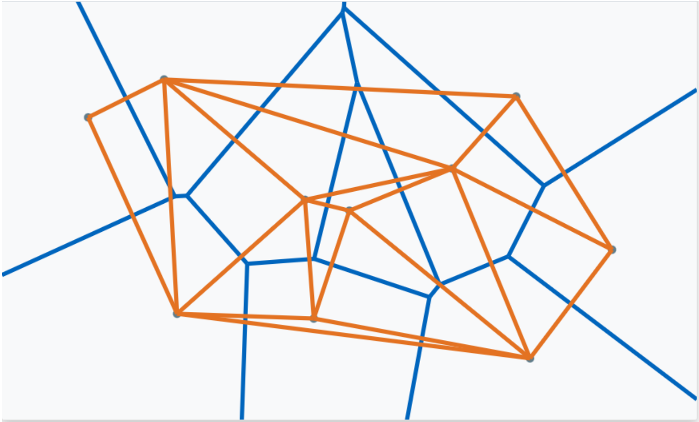
Voronoi diagram for a set of points



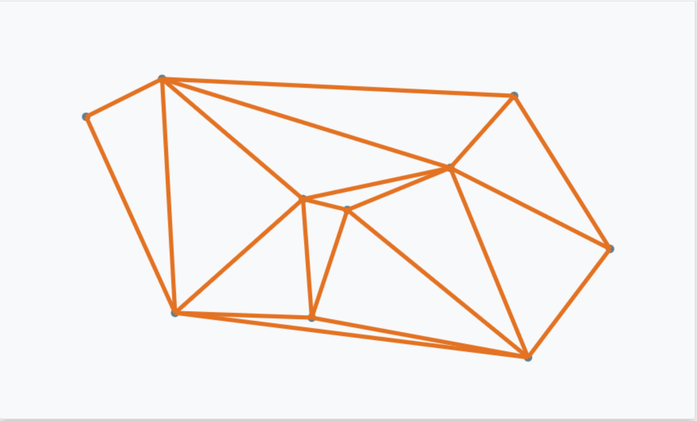
Voronoi diagram to Delaunay graph



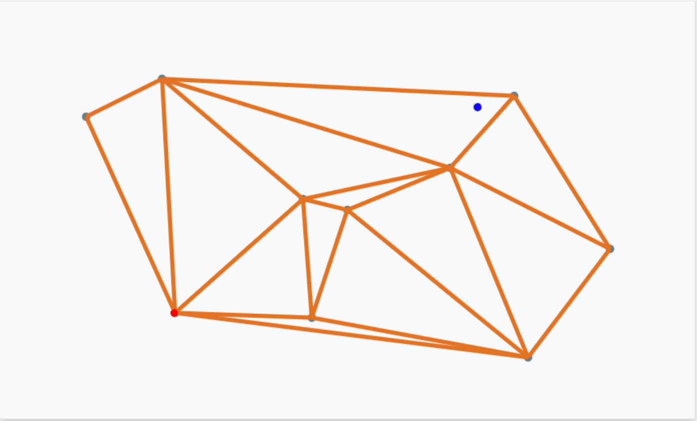
Voronoi diagram to Delaunay graph



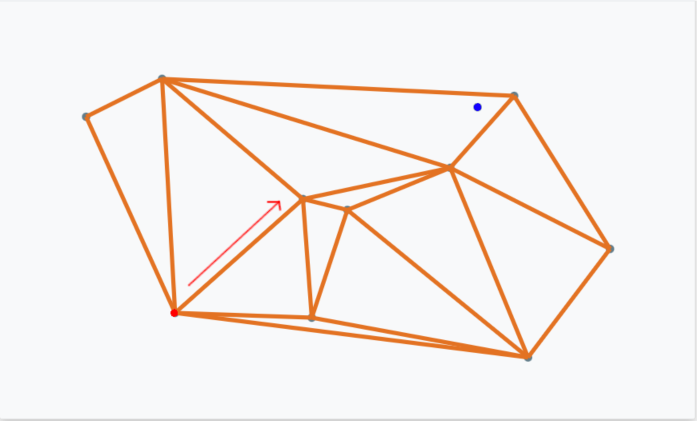
Delaunay graph



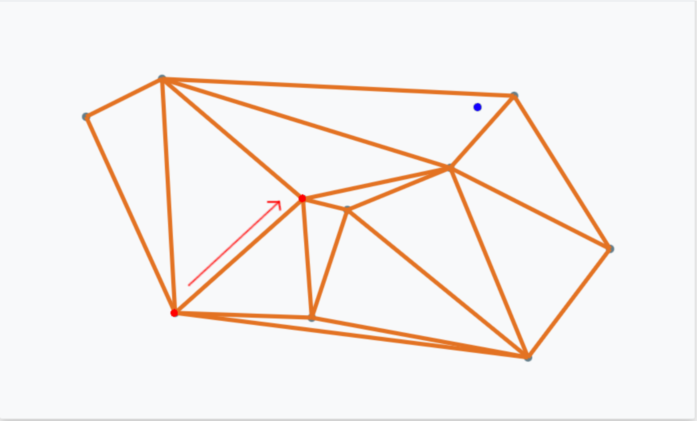
Greedy NN search start - Query and entry point



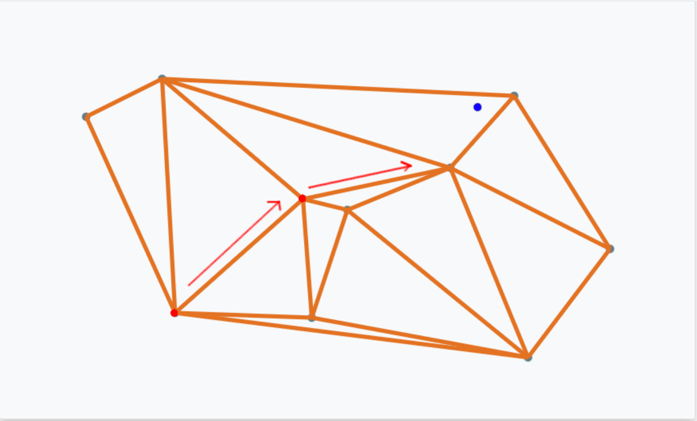
Greedy NN search - iteration



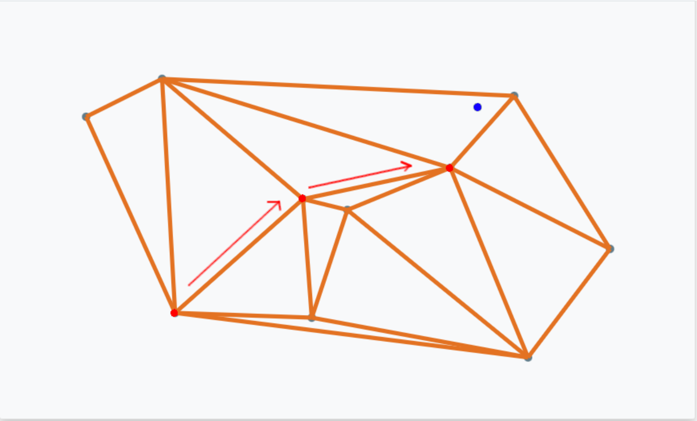
Greedy NN search - iteration



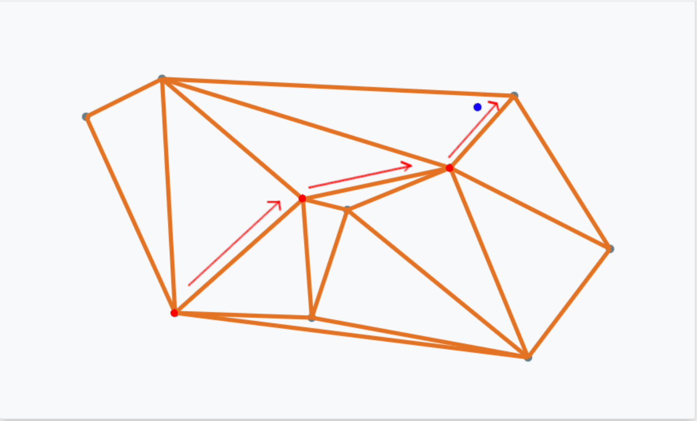
Greedy NN search - iteration



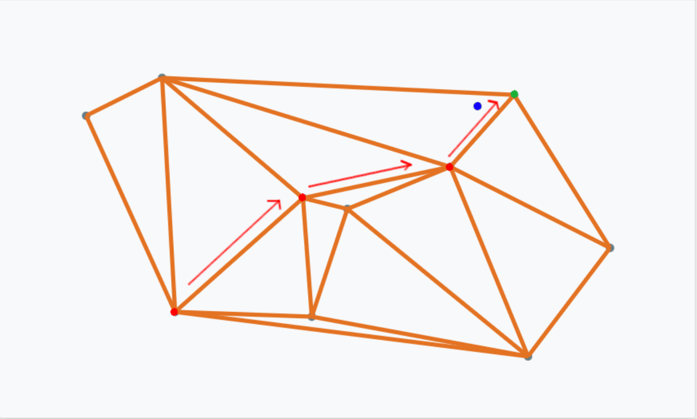
Greedy NN search - iteration



Greedy NN search - iteration



Greedy NN search done!



Asymmetric distance functions

Why?

- Can help us the direction of the prior art
- User might be interested also in finding infringing patents
- Currently we can't model this asymmetry

Asymmetric distance functions

How?

- We can learn a distribution for a document instead of just a single vector
- Model prior art relation as KL divergence of distributions

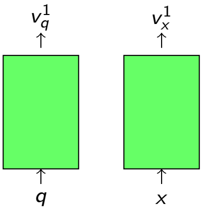
Distance functions for metadata

Why?

- We can do soft filtering (by country, patent class etc.)
- Can be useful if match is not found by strict filters

Learning multiple distance functions - naive way

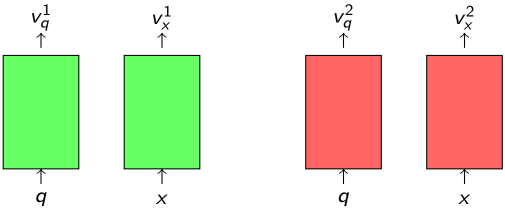
$$d_1(q,x) = v_q^1 \cdot v_x^1$$



Learning multiple distance functions - naive way

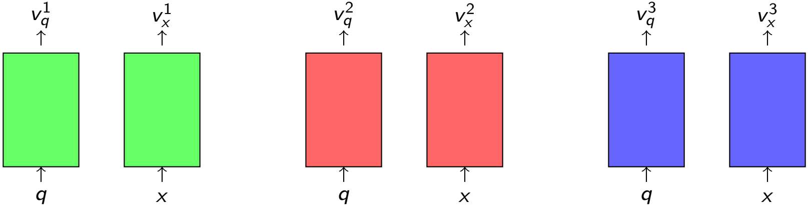
$d_1(q, x) = v_q^1 \cdot v_x^1$

$d_2(q, x) = v_q^2 \cdot v_x^2$



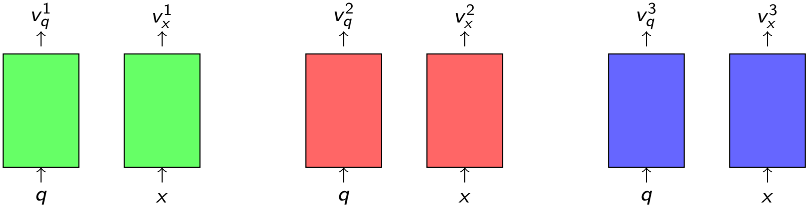
Learning multiple distance functions - naive way

$d_1(q, x) = v_q^1 \cdot v_x^1$ $d_2(q, x) = v_q^2 \cdot v_x^2$ $d_3(q, x) = v_q^3 \cdot v_x^3$



Learning multiple distance functions - naive way

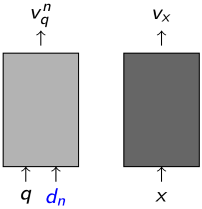
$d_1(q,x) = v_q^1 \cdot v_x^1$ $\quad$ $d_2(q,x) = v_q^2 \cdot v_x^2$ $\quad$ $d_3(q,x) = v_q^3 \cdot v_x^3$



Drawback: multiple embeddings of same document must be indexed!

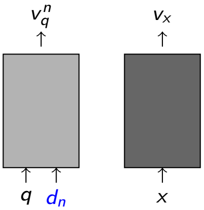
Learning multiple distance functions - efficient way

$$d_n(q, x) = v_q^n \cdot v_x$$



Learning multiple distance functions - efficient way

$$d_n(q, x) = v_q^n \cdot v_x$$



Only one meta-embedding per document is indexed!

Forward thinking

Forward thinking

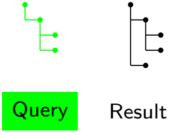
Why?

- We can train deeper models but keep batch size the same
- Training of deep models can take less wall clock time

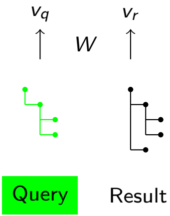
Forward thinking - paper

- Forward Thinking: Building and Training Neural Networks One Layer at a Time (Hettinger et al.) https://arxiv.org/abs/1706.02480

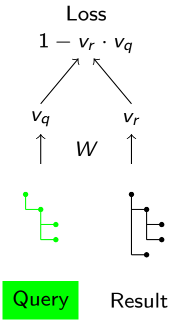
Current method - using gradients



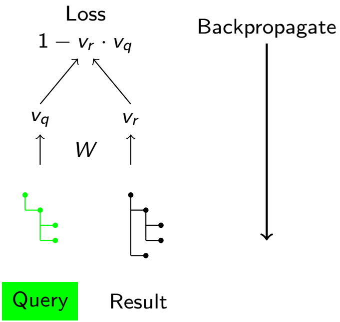
Current method - using gradients



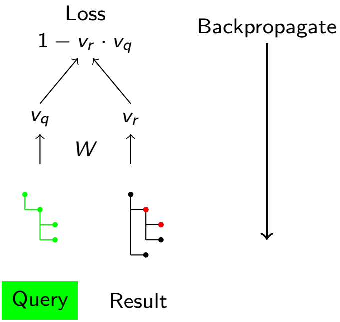
Current method - using gradients



Current method - using gradients



Current method - using gradients

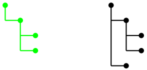


Nodes with highest gradient are considered most important

Drawbacks with using gradients

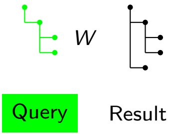
- Compute-intensive, since we need to do backwards pass
- Quality of explanations is not the best
  - Evaluating Recurrent Neural Network Explanations (Arras et al.) `https://arxiv.org/abs/1904.11829`

Comparing node embeddings



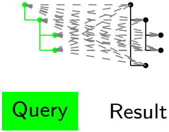
Query Result

Comparing node embeddings



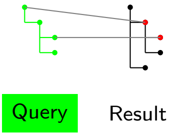
1. Embed graphs using model

Comparing node embeddings



1. Embed graphs using model
2. Compare each pair of node embeddings

Comparing node embeddings



1. Embed graphs using model
2. Compare each pair of node embeddings
3. Highlight most similar nodes

Comparing node embeddings

Why?

- Faster than using gradients (no backprop step needed)
- Might give more relevant explanations
- Can be useful for finding missing features

Summary

- We can give model more relevant data by using the citations and metadata more efficiently
- Learning distributions instead of just embeddings enables modelling asymmetry of prior art relations
- Might be possible to get better explanations by comparing node embeddings