Similarity searching

Filippo Geraci

May 14, 2019

Similarity search - Motivations

Scenario

- Text retrieval systems provide a *similarity search utility*, that allows users to find efficiently documents that are the *most similar* to the query.

Similarity searching

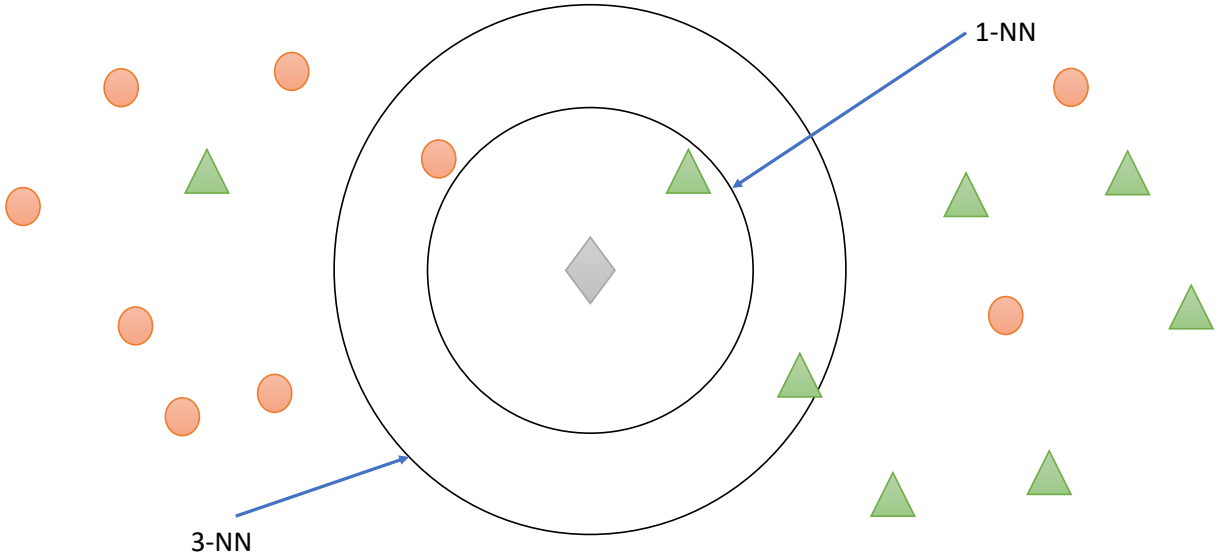
The problem

1. Find the k elements more similar to a given query
2. In absence of information all the pairwise distances must be computed

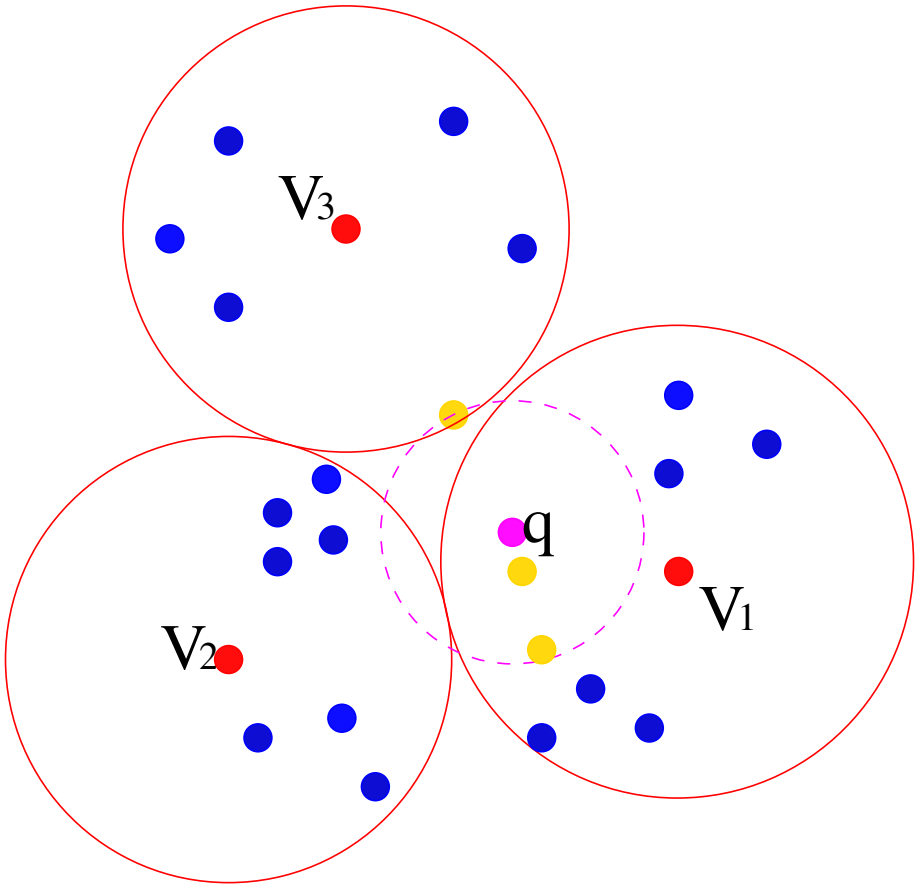
Caveats

1. The database can be typically huge.
2. A faster and approximate answer might be acceptable.
3. The system evolves in the time (new items can be added to the database).
4. A single distance computation can be expensive for certain data (i.e. texts)

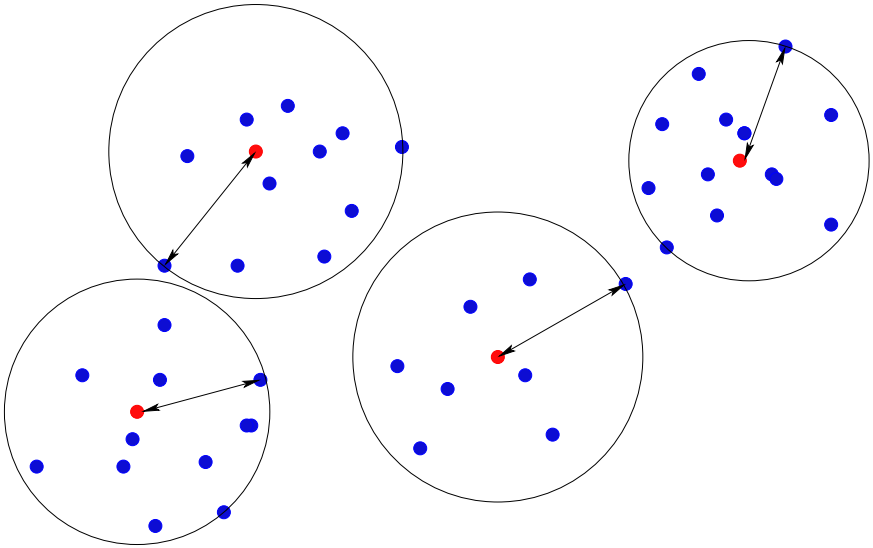
Similarity searching and classification



An approximate solution with clustering



Analogies with the k-center problem



- The closer a query is to a cluster center the more probable is that its nearest neighbors are in the cluster.
- Cluster radius bounds the distance among the query and its nearest neighbors in the cluster.