NEAREST NEIGHBORS PROBLEM

DJ O'Neil

Department of Computer Science and Engineering,
University of Minnesota

SYNONYMS

nearest neighbor, all-nearest-neighbors, all-k-nearest neighbors, reverse-nearest-neighbor-problem, reverse-k-nearest-neighbors aNN, akNN, rNN, rkNN, range query

DEFINITION

Given a set of n points and a query point, q , the nearest-neighbor problem is concerned with finding the point closest to the query point. Figure 1 shows an example of the nearest neighbor problem. On the left side is a set of $n = 10$ points in a two-dimensional space with a query point, q . The right shows the problem solution, s .

Figure 1: An example of a nearest-neighbor problem domain and solution.

The nearest-neighbor problem also includes the following problems:

- **k-nearest-neighbors (kNN):** Given a value $k \leq n$, kNN finds the k nearest objects to the query object. In most cases, the solution is the ordered k -nearest neighbors where the objects in the solution are ranked closest to farthest from the query point.
- **all-nearest-neighbors (aNN):** aNN is essentially NN applied to every point in the dataset.
- **all-k-nearest-neighbors (akNN):** akNN is kNN applied to every point in the dataset. Both akNN and aNN are usually used when NN queries will be applied to the data many times.
- **reverse-nearest-neighbor (rNN):** given a query point, q , rNN finds all points in the dataset such that q is their nearest neighbor.
- **reverse-k-nearest-neighbor (rkNN):** rkNN is similar to rNN except that it finds all points such that the query point, q , is in the set of their k -nearest-neighbors.

Within the GIS context, nearest neighbor problems are common in spatial queries where a user wants to know the closest objects, usually of some type, nearest to a query point. An example of this would be a query for finding what gas stations are near a particular address.

HISTORICAL BACKGROUND

The nearest-neighbor heuristic was developed as a greedy approach to approximating the traveling salesman problem (TSP). This problem dates as far back as the 1800s when related problems in graph theory were discussed by the Irish mathematician, Sir William Rowan Hamilton, and by the British mathematician, Thomas Penynghon. [The formal definition of the problem was first studied by Karl Menger at Harvard and Vienna in the 1930s. The TSP has gone on to be one of the classic graph problems in computer science. [3] More recently, the nearest neighbor problem has appeared in GIS, artificial intelligence, pattern recognition, clustering techniques, and outlier detection.

SCIENTIFIC FUNDAMENTALS

The nearest-neighbor algorithm is as follows:

Given:

- a set of **n** points.
- a query point, **q**.
- an initial minimum computed distance, **min_distance** = ∞ .

For each point, n_i , $i = 0, 1, \dots, n-1$:

- 1: Compute the distance between points **q** and **n_i** .
- 2: If the computed distance, **d**, is less than **min_distance**, then set **min_distance** = **d**

Solution: the value of **min_distance**.

The set of points can be defined in any space so long as a distance function is defined. Common distance functions in metric spaces are the Euclidean and Manhattan distances. The Euclidean distance is the distance found by determining the length of the straight line segment that joins the points. The Manhattan distance is (shortest) summed distance between the points using segments that are right angles to each other. The relationship between these two distance functions is that they represent the 2 paths between the points using a right triangle where the Euclidean distance is the hypotenuse and the Manhattan distance represents the legs.

The nearest-neighbor algorithm has two classical contexts. The first has to do with simply finding the nearest neighbor of some query point and the second uses neighbors as a simple classification technique.

Consider an example of the first type, such as finding the nearest gas station. Figure 2 shows a search from Google Maps for gas stations near the Metrodome. Here the **n** points consist of every geographical object that has an address. The query point is the Metrodome. In GIS applications, such as this, the data points are usually filtered to reduce the number of candidate points. This might include only considering points within a certain geographic region (e.g. Minneapolis, MN) and of a specific service type (e.g. Gas Station). The nearest search results are the nodes labeled **A**, **B**, **C**, and **D** with **A**

being the closest, then **B**, and so on. This problem is slightly different than the simple example illustrated in Figure 1. Because gas stations can only be reached by driving on the road, we cannot simply compute the Euclidean distance between the start point and candidate gas stations. In this example the distance function needs to consider multiple (reasonable) paths, constraints such as one-way streets, speed limits, and possibly other factors.

Figure 2: Finding the nearest gas station. (image from Google Maps. ©2006 Google – Map data ©2006 NAVTEQ™)

This example also demonstrates the k -nearest-nearest neighbors search. In this case, n could be all the gas stations in the world and k could be four. The solution to the problem is the gas stations identified by nodes **A**, **B**, **C**, and **D**.

Another common use of the nearest-neighbor problem is simple classification of a d -dimensional dataset. For example, suppose we wanted to predict which way someone would vote based on the following ($d = 5$):

- Age
- Height
- Weight
- Years of education
- Income level

Furthermore, suppose we took a sample of the population and collected this data along with their political affiliation; for simplicity we'll assume they can only be a Democrat or a Republican. Then, given the above data for an unclassified sample, the nearest-neighbor algorithm would use a distance function, f , to determine which 5-tuple in the classified dataset set is closest. The predicted classification of this new data point would be the same as the classification of its nearest neighbor.

A major problem with the simple nearest-neighbor algorithm is that it considers the entire set of n points for every execution. However, consider the Ann and Aknn problems where the same dataset is used n times. It would be beneficial to get some information about the dataset and the relationship between points in order to decrease the search space.

Figure 3. An example of a Voronoi Diagram.

One of the simpler methods is to use Voronoi Diagrams. Given a set of points, a Voronoi Diagram can be constructed such that every point in the region surrounding a data point is closest to that data point than any other. Figure 3 depicts a Voronoi Diagram. Consider the query point, q . To reduce the search space, one needs only to consider those points whose representative regions (dark gray) border the region of the query point (light gray). For values of k greater than the number of bordering regions, the algorithm could a subset of regions that border the union of the original candidate regions, and so on.

KEY APPLICATIONS

The nearest-neighbor problem and its variants are used in a variety of applications domains.

Spatial Network/Geographical Information Systems (GIS)

A spatial network is any network of spatial objects. GIS is an application of spatial networks that consist of points of various types of geographical data such as cities, parks, gas stations, etc... Spatial network problems, such as GIS, are different than classical Euclidean space queries because they need to consider spatial network distance. Spatial distance might be more concerned with time, whereas geometric distance exclusively considers Euclidean distance. Consider the nearest gas stations problem described

earlier: in addition to length traveled using roads, GIS time distances might also consider speed limits and traffic jams depending on the time of day. Because of these added metrics, the computed distance may not be the shortest. [9]

Due to increased storage space and processing power, spatial databases have to continued to grow in size and complexity. GIS applications, such as Google's Google Earth, contain detailed images and maps, extensive information on businesses and business profiles, and a huge address database. Because of the sheer volume of data to examine for nearest neighbor queries, new approaches to solving nearest neighbors have been proposed. These include approximation techniques that use reasonable assumptions to decrease the search space and novel data structures (k-d Trees, Range trees) designed with nearest-neighbors in mind.

Artificial Intelligence (AI)/Pattern Recognition

Pattern recognition is similar to the classification method discussed earlier. In AI, pattern recognition has been used in such problems as learning and vision by using example-based methods such as the nearest-neighbor. For machine learning, assuming the system has a large dataset of classified tuples, given a new data tuple, the nearest-neighbor algorithm can be applied on the classified data set. The new tuple will take on the classification of the nearest classified tuple. This method has been extended to use knn where the k-nearest tuples are located and the new tuple takes on the dominating classification of the k nearest tuples (ties broken by increasing or decreasing k). Similarly, this same method can be applied to images. For example, in [10], a hand pose was estimated using 100,000 hand images with known pose positions using the chamfer distance.

Clustering

The clustering problem is concerned with identifying associations between data points. Using the nearest-neighbors algorithm and a maximum distance threshold, the nearest neighbors can be incrementally identified from closest to farthest until the threshold has been reached. Clustering is used in many disciplines, ranging from marketing to geography to census data. [11]

Outlier Detection

An outlier is a deviating point that falls outside the overall pattern of a data set. In clustering and statistical problems, outliers are considered noisy data and are often removed. However, in data mining, an outlier may indicate an important anomaly. This has been used in intrusion detection systems where a normal pattern of facility use is established and outliers are classified as possible intruders. Aknn has been used in this problem by first identifying clusters in a controlled dataset. Then, given a new data point, p , the knn are computed. If computed knn fall into more than T clusters, then the data point is considered to be an outlier. [12]

Statistics

Because nearest neighbors can provide a quick and intuitive method to classify new data, the method has been applied to statistical methods. A primitive approach might simply

find the assign a new data point the classification of the nearest classified point. However, using kNN, confidence intervals and hypothesis tests for classifications can be applied to new data points by finding the set of k classified points nearest to the new data point.

Business

In addition to some of the applications listed above, such as AI and statistics, nearest neighbors approaches have been applied to many common business problems. For example, suppose a business would like to open up a new branch at one of several locations. A rkNN search could be applied to the candidate locations to predict which would have the largest customer base using the demographics of its surrounding geographic locations. This can also be extended to predict how many service requests a particular business location might experience.

FUTURE DIRECTIONS

As a brute force technique, the traditional nearest-neighbor algorithm isn't very useful with large datasets; especially if the dataset is going to be reused. When considering how much information is contained in many GIS applications, it becomes more apparent that searching the entire search space for a objects near the query point would be extremely costly and ineffective. Current research in nearest-neighbor algorithms is concerned with determining ways to approximate the solution or use nearest-neighbors as an approximation for other problems. For example, in [8], highly dimensional data points are first restructured into a balanced box-decomposition (BBD) tree where points are classified by hyper-rectangles with a bounded width vs. height ratio. Similar to an R-tree, the BBD tree can be used to quickly decrease the search space. Knn has also been used in GIS applications to estimate the most likely candidates of a spatial query. Finding the knn using Euclidean distance is straightforward and simple and saves more time than computing knn using complex spatial distances.

There has also been research in the appropriateness of the applying Euclidean and/or Manhattan distance functions in highly dimensional spaces. In [14] they showed that as dimensionality increases (10-15 dimensions), the distance to the nearest and farthest data points approach each other using traditional distance functions. Highly dimensional data also complicates the use of traditional spatial data structures. In [13] they observe that as dimensions increase, the overlap of minimum bounding boxes in R-trees increases significantly. They propose a new data structure, called the X-tree, that aims to keep data as hierarchical as possible to reduce overlap using the concept of supernodes.

CROSS REFERENCES

- Geographic Information System (GIS)
- Spatial Range Query
- Voronoi Diagram
- R-tree

RECOMMENDED READING

- [1] Anany Levitin. Introduction to The Design and Analysis of Algorithms. Addison Wesley. 2003.
- [2] The Algorithm Design Manual. “Nearest Neighbor Search”. Jun. 2, 1997.
- [3] Traveling Salesman Problem. “History of the TSP”. Jan. 2005.
- [4] Google. “Google Maps”. October 12, 2006. Search on “gas stations” near the “metro dome”.
- [5] Chavez, Edgar, Karina Figueroa, et al... “A Fast Algorithm for the All k Nearest Neighbors Problem in General Metric Spaces”
- [6] Das, Abhinandan, Johannes Gehrke, et al... “Approximation techniques for spatial data”. Proceedings of the 2004 ACM SIGMOD international conference on Management of data. Paris, France. SESSION: Research sessions: spatial data. Pgs. 695-706. 2004.
- [7] Tunkelang, Daniel. “Making the Nearest Neighbor Meaningful”. SIAM WORKshop on Clustering High Dimensional Data and its Applications. 2002.
- [8] Arya, Sunil. “An optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions”. Journal of the ACM (JACM), 1998. 1998.
- [9] Ku, Wei-Shinn, Haojun Wang, et al... “ANNATO: Adaptive Nearest Neighbor Queries in Travel Time Networks”. 7th International Conference on Mobile Data Management 2006. May 10-12, 2006. pg. 50.
- [10] Athitos, Vassilis. “Approximate Nearest Neighbor Retrieval Using Euclidean Embeddings”. Approximate Nearest Neighbors Methods for Learning and Vision. Dec 13, 2003.
- [11] Painho, Marco, Fernando Bacao. “Using Genetic Algorithms in Clustering Problems”. GeoComputation 2000. 2000.
- [12] Hautamaki, Ville, Ismo Karkkainen, et al... “Outlier Detection Using k-Nearest Neighbor Graph”. Proceedings of the 17th International Conference on Pattern Recognition, 2004. Volume 3. Aug. 23-26, 2004. pp.430-433.
- [13] Berchtold, Stefan, Daniel A. Keim, Hans-Peter Kriegel. “The X-tree: An Index Structure for High-Dimensional Data”. Proceedings of the 22nd International Conference on Very Large Databases. 1996.
- [14] Beyer, Kevin, Jonathon Goldstein, Raghu Ramakrishnan, Uri Shaft. “When is ‘Nearest Neighbor’ Meaningful?”. Lecture Notes in Computer Science. 1999.

NEAREST NEIGHBORS PROBLEM

DJ O'Neil

Department of Computer Science and Engineering,
University of Minnesota

SYNONYMS

nearest neighbor, all-nearest-neighbor, all-k-nearest neighbor, Ann, Aknn, range query

DEFINITION

Given a set of n points and a query point, q , the nearest-neighbor problem is concerned with finding the point closest to the query point. Figure 1 shows an example of the nearest neighbor problem. On the left side is a set of $n = 10$ points in a two-dimensional space with a query point, q . The right shows the problem solution, s .

The nearest-neighbor problem also includes the following problems:

- **k-nearest-neighbors (kNN):** Given a value $k \leq n$, kNN finds the k nearest objects to the query object. In most cases, the solution is the ordered k -nearest neighbors where the objects in the solution are ranked closest to farthest from the query point.
- **all-nearest-neighbors (aNN):** aNN is essentially NN applied to every point in the dataset.
- **all-k-nearest-neighbors (akNN):** akNN is kNN applied to every point in the dataset. Both akNN and aNN are usually used when NN queries will be applied to the data many times.
- **reverse-nearest-neighbor (rNN):** given a query point, q , rNN finds all points in the dataset such that q is their nearest neighbor.
- **reverse-k-nearest-neighbor (rkNN):** rkNN is similar to rNN except that it finds all points such that the query point, q , is in the set of their k -nearest-neighbors.

MAIN TEXT

The nearest-neighbor algorithm is as follows:

Given:

- a set of n points.
- a query point, q .
- an initial minimum computed distance, $\text{min_distance} = \infty$.

For each point, n_i , $i = 0, 1, \dots, n-1$:

- 1: Compute the distance between points q and n_i .
- 2: If the computed distance, d , is less than min_distance , then set $\text{min_distance} = d$

Solution: the value of min_distance .

The set of points can be defined in any space so long as a distance function is defined.

The nearest-neighbor algorithm has two classical contexts. The first has to do with simply finding the nearest neighbor of some query point and the second uses neighbors as a simple classification technique. Figure 2 depicts an example of the first use where

someone wants to know which gas station is nearest to the Metrodome. Here the **n** points consists of every geographical object that has an address. The query point is the Metrodome. A simple classification technique using nearest-neighbors would classify a new data point based on the classification of the nearest surrounding points.

CROSS REFERENCES

- Geographic Information System (GIS)
- Spatial Range Query
- Voronoi Diagram