Approximate Nearest Neighbors for Structured Data

Advisor: Ioannis Emiris (http://www.di.uoa.gr/∼emiris/index-eng.html)
Lab of Geometric & Algebraic Algorithms (http://erga.di.uoa.gr/), University of Athens, Greece
Duration: 2 to 3 months

Topic. Given a pointset in general dimension, a fundamental question is to find the approximately nearest point (ANN) in the set, for any query point. It is encountered in many critical areas, including optimization, searching and machine learning, in a variety of applications ranging from bioinformatics and GIS to websearch. There has been a number of sophisticated methods, achieving logarithmic query time in the number of points, but typically they do not exploit structure in the input, although such structure is manifest in many applications. Exploiting structure offers today an important means for going beyond the state of the art.

We examine ANN when data points lie in low dimensional manifolds. We start with points assumed to be almost aligned on an unknown (small) number of unknown lines. Different scenarios shall be investigated, depending on the point distribution on the lines and the line distribution in space. Under certain assumptions, by storing the input in data structures based on kd-trees, see e.g. [1], query time is expected to be logarithmic in the number of lines, instead of in the number of points, when dimension is not too large. A key step is to reduce finding the nearest line to another ANN problem [2].

During the internship, a query algorithm and a data structure for the input will be developed in C++, with the goal of implementing them in CGAL [3]. We aim at obtaining good performance in practice, in important application areas. If time permits, we may examine other types of structure, e.g. curves. *For students in Maths-Info, the project would include a probabilistic error analysis and an asymptotic complexity analysis, and the preparation of a report on them.*

The work is inscribed in European project "Computational Geometric Learning" (http://cglearning.eu/) whose goal is to design data structures and algorithms for complex problems in high dimensions by exploiting properties of the input.

References

[1] Arya, S., Malamatos, T., Mount, D.M.: Space-time tradeoffs for approximate nearest neighbor searching. J. ACM 57(1) (2009)

[2] Basri, R., Hassner, T., Zelnik-Manor, L.: Approximate nearest subspace search. IEEE Trans. Pattern Anal. Mach. Intell. 33(2), 266–278 (2011)

[3] CGAL: Computational Geometry Algorithms Library. `http://www.cgal.org.`