Semester Report Laura Heinrich-Litan

Supervisor:	Prof. Dr. Helmut Alt
Field of Research:	Nearest Neighbor Search
Topic	Exact Nearest Neighbor Search in High Dimensions
PhD Student	in the program since April 1999

Field of Research

The *nearest neighbor problem* is the problem of constructing an efficient data structure storing a set P of n points in $\mathbb{R}^d$ for answering *nearest neighbor queries*. In a nearest neighbor query some point $q \in \mathbb{R}^d$ is specified and the (some) point in P closest to q has to be determined.

In many applications, the dimension d of the search space is quite high and can reach several hundreds or even several thousands. Therefore, running times and storage requirements exponential in d are prohibitive in these cases. All query algorithms are in fact competing with the *brute-force* method which requires no preprocessing, storage only the set P, and has a query time of $O(nd)$ for all L_p-distances, $1 \leq p \leq \infty$.

Results

My work concentrates on exact nearest neighbor search with respect to the L_∞-distance. The results can be easily extended to other box-distances (where the unit ball is a rectangular box). In the previous semesters I considered query methods which do not need preprocessing. These algorithms (they are called *cube-methods*) have an average runtime of $\Theta(nd/ln\ n + n)$ assuming that the set P of n points is drawn randomly from the unit cube $[0,1]^d$ under uniform distribution; thereby improving the brute-force method essentially by a factor of $\Theta(1/ln\ n)$. I generalized the methods and their analysis to other "well-behaved" probability distributions and to the important problem of finding the k nearest neighbors to a query point. The k nearest neighbors from P to some query point $q \in [0,1]^d$ can be found with an expected asymptotic runtime of $O\left(\frac{nd}{\ln(n/k)} + n + dk + d\log d\right)$. The results are summarized in [1].

Further results investigate a method which provides tradeoffs between the space complexity of the data structure and the time complexity of the query algorithm. The average runtime of the query algorithm assuming that the set

P is drawn randomly from the unit cube $[0,1]^d$ under uniform distribution is essentially $\Theta(d \cdot (1 + \frac{\sqrt[d]{n}}{m})^d)$ where $m \geq 2$ is a parameter. The expected storage required by the data structure is $\Theta(d(m + \sqrt[d]{n})^d)$. The results are summarized in [2].

Using a partition of the point set into monotone subsequences, which is done in the preprocessing phase, the *cube-methods* can be improved by a speedup factor of $\Omega(n^{\frac{1}{\sqrt{d}}})$, i.e. the query time is $O(\frac{dn^{1-\frac{1}{\sqrt{d}}}}{\ln n} + n^{1-\frac{1}{\sqrt{d}}})$ assuming that the set P is drawn randomly from the unit cube $[0,1]^d$ under uniform distribution. Further improvements which I investigated this semester provide a method with query time $O\left(n \ln\left(\frac{d}{\ln n}\right) + n\right)$ and linear storage requirement. The method has also an efficient external-memory variant.

Activities

- I took part in the course *Algorithmen für Fortgeschrittene*, held by Helmut Alt.
- Participation at the lectures and colloquia of the graduate program.
- Talk in the colloquim of the graduate program on January 29, 2001.
- Participation at ADIMMO 2001, March 15-16, 2001, Trier.
- Participation and talk at the Dagstuhl Seminar on Computational Geometry, March 18-23, 2001.
- Participation and talk at the 17th European Workshop on Computational Geometry, March 26-28, 2001, Berlin.
- Participation and talk at Workshop on Combinatorics, Geometry, and Computation, May 13-15, 2001, Ascona.
- Participation and talk at the 17th ACM Symposium on Computational Geometry, June 2001, Medford, Massachusets, USA.
- Referee for IEEE Transactions on computers

Preview

I am and will be working on finishing my thesis. As this is my last semester in the graduate program I would like to thank all those who contributed to the excellent research and educational conditions in the program. Special thanks are addressed to my supervisor Prof.Dr. Helmut Alt and to Bettina Felsner.

References

[1] H. Alt and L. Heinrich-Litan. Exact L_∞ Nearest Neighbor Search in High Dimensions. In *Proceedings of the 17th ACM Symposium on Computational Geometry*, June 2001.

[2] L. Heinrich-Litan. Time-Space tradeoffs for exact L_∞-Nearest-Neighbor-Search in High Dimensions. *17th European Workshop on Computational Geometry*, Berlin, March 2001.