COP 6731 Advance Database Systems

Project Proposal

Anmol Sureshkumar Panchal; UID: 4446829

Topic: Locality Sensitive Hashing (LSH)

Objective:

To study LSH and algorithms to improve its performance to process large datasets of information retrieval, data mining, image processing and so on.

Introduction:

A similar search problem like KNN queries are often aimed at collection of objects (e.g., documents, images) that are characterized by a collection of relevant features and represented as points in a high dimension attribute space. Given that those queries are in the form of points in the space, we are required to find the nearest (most similar) objects to the queries. The interesting and well explored case is d-dimensional Euclidean space. This problem is of major importance to a variety of applications like data compression, databases and data mining, information retrieval, image and video databases, data analysis. In these areas, the problem that usually arises is arrival of vast amount of data and of high dimensions, which we also call “curse of dimensionality”. When given a collection of points and a distance function such as hamming distance function or Euclidean distance function between them, the nearest search is one of the most important forms in similarity search. Nearest neighbor search has been applied in image processing, local match feature etc.

The key insight behind the technique of LSH is that the hash functions should map adjacent points to the same buckets with higher probability, but map points far from each other to the same buckets with lower probability. The basic idea of LSH is to hash the points from the database to ensure that the probability of collision is much higher for those points that are close to each other than for those that are far apart. LSH is an indexing technique that makes it possible to search efficiently for nearest neighbors amongst large collections of items, where each item is represented by a vector of some fixed dimension. The algorithm is approximate but offers probabilistic guarantees i.e. with the right parameter settings the results will rarely differ from doing a brute force search over your whole collection. The search time will certainly be different though: LSH is useful because the complexity of lookups becomes sublinear in the size of the collection. Now later in report we will see the techniques proposed in the three papers selected to present and will analyze and conduct the comparisons to see which method is more suitable for LSH implementation. The title of the papers are as follows:

Paper 1: - An Improved Algorithm for Locality-Sensitive Hashing

Paper 2: - Entropy based Locality-Sensitive Hashing

Paper 3: - Frequency Based Locality Sensitive Hashing

References:

Wu, T., & Miao, Z. (2016). An improved feature image matching algorithm based on Locality-Sensitive Hashing. 2016 IEEE 13th International Conference on Signal Processing (ICSP). doi:10.1109/icsp.2016.7877927

Wang, Q., Guo, Z., Liu, G., & Guo, J. (2012). Entropy based locality sensitive hashing. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp.2012.6288065

Ling, K., & Wu, G. (2011). Frequency Based Locality Sensitive Hashing. 2011 International Conference on Multimedia Technology. doi:10.1109/icmt.2011.6002015a