論文の内容の要旨

A Study of Fast Similarity Search Techniques in Metric Spaces

（メトリック空間における類似検索の高速化に関する研究）

氏名　倉沢　央

The purpose of my research is to reduce the query execution cost of a similarity search in Metric spaces. Although a large number of studies have been made on indexing techniques for similarity searches, only a few exploit the data distribution in a Metric space to indexing. My goal is to develop a new partitioning scheme based on the data distribution that can prune objects while the search more effectively.

Finding similar objects in a large dataset is a fundamental process in various applications, such as record linkage, and GIS. The reduction of the query execution cost of a similarity search can speed up these applications. Similarity searches based on a Metric space can be applied to all types of data whose distance obey Metric space postulates such as the triangle inequality. Therefore, Metric space indexes are very useful for applications that deal with huge amounts of vectors, strings, graphs, sets, and so on.

Similarity search indexes are used for pruning objects dissimilar to a query and reduce the search cost, such as the distance computations and the disk accesses. Most indexing schemes use pivots, which are reference objects. They recursively divide a region into subregions by using pivots and construct a tree structure index.

They prune some of the subregions by using the triangle inequality while searching. That is, the methods of selecting pivots and dividing up the space by using these pivots determine the index structure and pruning performance.

Pivot selection methods proposed in early researches were based on heuristics. They asserted that good pivots should be outliners of the space, because the distances from a pivot to each object vary and the pivot easily classifies the objects. They used simple statistical features such as the mean and the variance for selecting pivots. However, their choices of pivots are only a little better than random selection. Several proposed selection mechanisms exploit data distribution by clustering techniques. These methods set cluster center as the pivots and divide the space on the basis of the distances from the pivots. Although they can effectively classify dense regions, they only work well on particular distribution patterns. A cluster may be separated into multiple regions by a pivot, because their partitioning boundaries are based on cluster centers rather than cluster shapes. I focus on the problem in the existing indexes that they don't consider the partitioning boundary of the pivot and cannot select a good pivot for pruning.

I developed two novel methods for pivot partitioning called Maximal Metric Margin Partitioning (MMMP) and Pivot Capacity Tree (PCTree). The MMMP firstly extracts the data distribution pattern, especially for the boundaries of clusters. Then, it selects a pivot and its partitioning distance based on the shapes of data clusters. The partitioning boundary of the MMMP is at maximum distances from the neighbor cluster edges. The MMMP is good for dealing with clustered data. On the other hand, the PCTree considers the index tree balancing as well as the data distribution. The PCTree chooses a pivot based on both the balance of the subregions partitioned by a pivot and the estimated effectiveness of the search pruning by the pivot. As a result, PCTree automatically optimizes the index structure according to the data distribution. The PCTree improves the tree imbalance of the MMMP. These indexing methods successfully reduce the similarity search cost.

The main contribution is that I have developed a new pivot selection approach

called ``the data distribution-based approach", which is based on the data distribution. To my best knowledge, this is the first study to exploit the cluster shapes and the tree balancing for the pivot selection.

I show the efficiency of these methods empirically through the results from several experiments where I compared the methods with several Metric space indexes. In this thesis, I firstly introduce the basic techniques of similarity search, and then show my methods and experimental results.