学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　陸　可鏡

学　位　論　文　題　名

Query-Aware Locality Sensitive Hashing for Similarity Search Problems
（類似検索問題に対するクエリ・アウェア局所鋭敏ハッシング）

Similarity search is an old but fundamental research topic which has various applications in database, data mining, information retrieval and machine learning. Although the goal of similarity search, that is, finding most similar points of issued queries in the dataset given a specified measure, is quite straightforward, it is very challenging to solve this problem efficiently due to the following two reasons. (1) In recent years, as the data size increases rapidly, we need to deal with up to billion-scale datasets, on which many traditional techniques lose the effectiveness. (2) Due to the phenomenon called the curse of dimensionality, the distances among data points become much closer as the dimension increases, which makes most of tree structures perform even worse than the linear scan. In order to overcome these two obstacles, many researchers have devoted to find more efficient techniques in the past two decades and proposed various algorithms.

Although these algorithms vary much, the following two performance metrics always attract particular attention in their designs: one is the accuracy, which is often indicated by the percentage of true points which are successfully found, also called, the recall rate, and the other one is the efficiency which is often indicated by the running time of searching. In order to make the accuracy more controllable, researchers have developed some techniques with theoretical guarantees to satisfy specified success probabilities. Among these techniques, Locality Sensitive Hashing (LSH) draws a particular interest due to its attractive query performance and robust probability guarantee. The basic idea of LSH is to build a group of projected vectors and select the most promising candidates only from the projection information of those vectors. Such an idea becomes the starting point of the research work introduced in this paper.

Since the solutions of similarity search problems vary highly over different spaces and different measures, in this paper, we particularly focus on the most important two situations: the Euclidean space equipped with ℓ_2 metric and the inner product space. In practice, many applications are intimately related to either of these two situations. Accordingly, this thesis is divided into two major parts.

In the first part, we focus on the ℓ_2 metric and aim to solve approximate nearest neighbor search problems. For this purpose, we propose two disk-based LSH variants called VHP and R2LSH, which are highly inspired by the concept of query-aware search window. By exploiting more accurate projection information of the generated one-dimensional projected vectors, these two methods build more effective query-centric search regions in the projected spaces. Specifically, R2LSH builds multiple query-centric balls in a group of two-dimensional projected subspaces, while VHP utilizes the information of one-dimensional distances between the query and data points on every projected vector. Both of these two methods can prune false points more accurately than the existing query-aware LSH

variants.

In the second part, we focus on the inner product space and solve the maximum inner product search problem. For this purpose, we propose a query-aware LSH variant called AdaLSH which is built on a multi-ring structure. The basic idea of AdaLSH is that, based on different norms of data points, we can control adaptively the width of the query-aware search windows such that those points having larger norms can be examined more carefully to ensure the accuracy. In order to realize this idea, we design a multi-round search strategy such that query-aware search windows in different rings extend at different paces, which not only achieves the goal mentioned above, but also significantly decreases the total searching cost.

Since all three proposed methods are based on LSH, all of the proposed algorithms possess probability guarantees in accuracy. Extensive experiments confirm the superiority of these methods over existing state-of-the-art methods. In particular, R2LSH and VHP scale up to billion-scale datasets, because they are disk-based.