学位報告４

報告番号	※甲　　　　第　　　　号

主論文の要旨

論文題目　Query Processing over Probabilistic Data with Gaussian Distributions
（ガウス分布に基づく確率的データに対する問合せ処理）

氏　　名　董　婷婷（DONG Tingting）

論文内容の要旨

The field of uncertain data management has received extensive attention of researchers due to the increasing demand for managing uncertain data in a large variety of real-world applications such as sensor networks, location-based services, monitoring and surveillance. Uncertainty can occur for different reasons, including measurement errors and noises in sensors, privacy-preserving transformations of sensitive records and the limited confidence in the output of predictive models.

Managing uncertainty involves modeling, representing, querying, and indexing uncertain data. Answering queries over uncertain databases poses more challenges than that over traditional databases because managing uncertainty usually means costly probability computations. Hence, it is crucial to develop efficient solutions when managing uncertain data. In this thesis, we model uncertainty probabilistically and represent each uncertain object in the database using a Gaussian distribution, which is a typical probability distribution widely used in statistics, pattern recognition, and machine learning. We consider the following three types of queries or searches over probabilistic data with Gaussian distributions.

First, we study the probabilistic range query, which is an important query in the field of uncertain data management. A probabilistic range query returns objects in the database that exist within a specified range from the query object with probabilities no less than a given probability threshold. The query object can be either a certain point or an uncertain object represented by a Gaussian distribution.

We propose several effective filtering techniques by analyzing the properties of Gaussian distribution. The proposed filtering techniques can significantly reduce the number of objects that need to be verified by expensive probability computation. In this way, we can avoid unnecessary computations and hence save cost. To support efficient query processing, we further propose a novel indexing method to enable filtering unpromising objects by groups rather than individually. We develop the indexing method by extending the existing R-tree and improve it based on our analysis of Gaussian distribution. This indexing method can effectively organize objects and greatly enhance the performance of query processing. Extensive experiments on real datasets demonstrate the efficiency and effectiveness of our proposed approach.

Second, we investigate the nearest neighbor search. As one of the commonest queries over location information, the distance-based nearest neighbor search, which finds closest objects to a given query point, has extensive applications in many areas. There have been considerable efforts to extend nearest neighbor search over traditional location information to uncertain location information. An example is the expected distance, which defines the distance over uncertain location information. Following this trend, we represent uncertain locations using Gaussian distributions and assume that the closeness between each Gaussian object and the query point is measured by their expected distance. Under this setting, we consider the problem of *k*-expected nearest neighbor search over Gaussian objects. The result objects are ones that have the top-*k* smallest expected distances to the query point.

We analyze properties of expected distance on Gaussian distribution mathematically and derive the lower bound and upper bound of the distance. Based on our analysis, we propose three novel approaches to efficiently solve this problem. The proposed approaches can prune unpromising objects whose lower bound distances are larger than upper bound or expected distances of candidate objects without computing their actual expected distances. We only compute exact expected distances for candidate objects and finally return the top-*k* smallest ones. To further improve the performance, we utilize R-tree to index objects and their lower bound distances and upper bound distances. The proposed approaches can effectively reduce the number of exact distance computation which is rather expensive. The efficiency and effectiveness of our approaches are demonstrated through extensive experiments.

Finally, we explore the problem of similarity search, which is a crucial task in many real-world applications such as multimedia databases, data mining, and bioinformatics. In this work, we investigate similarity search on uncertain data represented by Gaussian distributions. The query object is also represented by a Gaussian distribution. By employing Kullback-Leibler divergence (KL-divergence) to measure the similarity between two Gaussian distributions, our goal is to search a database for the top-*k* Gaussian distributions similar to a given query Gaussian distribution. Especially, we consider non-correlated Gaussian distributions, where there are no correlations between dimensions and their covariance matrices are diagonal.

To support query processing, we propose two types of novel approaches utilizing the notions of rank aggregation and skyline queries. The first type presorts all objects in the database on their attributes and computes result objects by merging candidates from each presorted list. The second one transforms the problem to the computation of dynamic skyline queries. We extend and modify the branch-and-bound skyline (BBS) algorithm, which is proposed to answer skyline queries, and develop a novel algorithm to solve this problem. We demonstrate the efficiency and effectiveness of our approaches through a comprehensive experimental performance study.

In general, we provide a comprehensive view of managing probabilistic data with Gaussian distributions. We believe that our contributions are multi-dimensional and will be extended over time. First of all, we think our mathematical analyses of probabilistic range query, nearest neighbor search, and similarity search over Gaussian objects will benefit not only the existing related applications, but also potential studies of other applications over data represented by Gaussian distributions. Furthermore, our proposed non-trivial algorithms and indexing structures for efficient query processing, will not only enrich user experience by speed in the real world, but also provide valuable insights and references for developing solutions to other problems. Last but not least, we have conducted extensive experimental evaluations on the performance of our proposed solutions.

学位関係

38字×23行