PROCESSING INCOMPLETE *K* NEAREST NEIGHBOR SEARCH

ABSTRACT

Given a set *S* of multi-dimensional objects and a query object *q*, a *k* nearest neighbor (*k*NN) query finds from *S* the *k* closest objects to *q*. This query is a fundamental problem in database, data mining, and information retrieval research. It plays an important role in a wide spectrum of real applications such as image recognition and location-based services. However, due to the failure of data transmission devices, improper storage, and accidental loss, *incomplete* data exists widely in those applications, where some dimensional values of data items are *missing*. In this paper, we systematically study *incomplete k nearest neighbor* (I*k*NN) *search*, which aims at the *k*NN query for incomplete data. We formalize this problem, and propose an efficient LP algorithm using our newly developed index to support *exact* I*k*NN retrieval, with the help of two pruning heuristics, i.e., *value pruning* and *partial distance pruning*. Furthermore, we propose an approximate algorithm, namely HA, to support approximate I*k*NN search with improved search efficiency and guaranteed error bound. Extensive experiments using both real and synthetic data sets demonstrate the effectiveness of newly designed indexes and pruning heuristics, as well as the performance of our presented algorithms under a variety of experimental settings.

EXISTING SYSTEM:

Many efforts have recently been made in terms of incomplete data, including the methodologies on incomplete data indexing incomplete data and querying incomplete data. To the best of our knowledge, this is the first attempt on studying I*k*NN retrieval systematically. Furthermore, the most related work is the probabilistic similarity search on uncertain data. Nonetheless, due to the differences between uncertain data and incomplete data mentioned earlier, existing probabilistic similarity search algorithms (as to be surveyed in Section II.A) cannot be utilized directly to tackle I*k*NN queries. On the other hand, traditional similarity search algorithms for complete data cannot also be applied to answer I*k*NN retrieval because of the missing values.

DISADVANTAGES:

- *Incompleteness* is a common problem data sets
- Incompleteness refers to the case where data has error or certain information is missing
- we can simply perform all the analysis tasks based on complete data sets by discarding all the incomplete data, the output might be biased or inaccurate.

PROPOSED SYSTEM:

We identify and formulate the I*k*NN query, and introduce a method to measure the closeness between objects even with incomplete data values. We present a novel L B index structure with the support of lattice and bucket structures, and develop an exact LP algorithm based on the L B index to solve I*k*NN search, using efficient pruning heuristics. We propose an efficient approximate HA algorithm based on the HIT index to handle I*k*NN retrieval efficiently under the theoretical precision bound. We conduct extensive experiments with both real and synthetic datasets to verify the effectiveness of L B and HIT indexes and our presented heuristics, and the efficiency of our proposed algorithms under a variety of settings.

ADVANTAGE OF PROPOSED SYSTEM:

- The effectiveness of our developed indexes and our presented heuristics, and the performance of our proposed algorithms.
- To minimize the number of node accesses
- We develop the PV-index which stores the minimum bounding rectangles (MBRs) in a systematic manner in order to efficiently answer the probabilistic NN query.

SYSTEM SPECIFICATION

Hardware Requirements:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Sony.
- Ram : 512 Mb.

Software Requirements:

- Operating system : Windows 7.
- Coding Language : ASP.Net with C#
- Data Base : SQL Server 2005.