【主論文の要旨・記入例】

学位報告 4

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Spatio-Temporal Data Warehousing for Exploratory Analysis of Scientific Data
（科学データの探索的分析のための時空間データウェアハウスに関する研究）

氏　　名　　趙　菁

論 文 内 容 の 要 旨

In scientific fields, the amount of data has been increasing enormously owing to the development of information technologies, high performance computers, and high capacity storage. As one of the common data types of big data, *spatio-temporal data* has been widely applied in various domains such as mobile applications and scientific research. For instance, disaster simulations addressing a particular spatial area and calculate temporal variations in the field after disasters are conducted for the purpose of predictions, decision making, etc. Consequently, managing and analysis of large-scale scientific spaito-temporal data has been in great demand.

In this thesis, we study the *data warehousing* techniques that enabling massive data storage and data exploration for sophisticated analytic processing of large amount of data. Researches on data warehousing involve data modeling, basic operators like data join and aggregation, advanced operators support for exploratory analysis, etc. We focus on the above-mentioned aspects and mainly conduct the following three studies.

First, we propose a simulation data warehouse-approach for interactive analysis of massive simulation data. The objectives of this work include integrating different simulation data sets, as well as enabling exploratory analysis of multiple accumulated simulation data with high-speed response by data preprocessing. We develop a prototype system architecture consists of data storage based on a *multi-dimensional data cube*, as well as an analysis interface to enable basic operators such as drill-down and roll-up for interactive analysis.

In addition, we show the usability of the proposed prototype system by a case example using disaster simulation data.

Second, we study advanced operators for exploratory analysis by data summarization techniques, i.e., *histograms*. One of the challenges of exploration-based analysis of spatio-temporal data is the semantic meanings of different granularities such as one day, one hour and one minute on the time dimension, which leads to the exhausting exploration by basic operators. Therefore, more advanced operators such as effective summarization and visualization that navigates users to find interesting knowledge are required.

We study the problem of constructing histograms (i.e., the *spatial V-optimal histogram*) that summarize the data distribution of a specific spatial area during a time interval. We propose exact and approximate algorithms, as well as a heuristic algorithm for efficient construction of hierarchical histograms. As scientific data is usually represented as 2-D (or 3-D) array structure, *array*-based representation is more appropriate for the representation. The proposed methods are implemented on the state-of-the-art array DBMS, SciDB, which supports efficient scientific computing and analysis of spatio-temporal array data. In addition, we conduct extensive experiments on massive simulation data, real taxi data as well as synthetic data with different distributions, to verify the effectiveness and the efficiency of the proposed methods.

Finally, we study the *difference* analysis of spaito-temporal data on a data warehouse, to detect temporal variations as well as differences between observation datasets with different parameters or conditions. We propose a general framework for constructing histograms on SciDB, as well as both optimal and heuristic histogram construction algorithms based on the structure of quadtree. We conduct experiments on massive simulation data to evaluate the performance of the algorithms, with a case study of the proposed difference operator.

In general, we provide data warehousing techniques including system prototypes, data modeling and advance operators to manage and analyze large-scale spatio-temporal scientific data. In order to enable efficient data visualization and effective data analysis, our proposed methods involve data preprocessing, approximate and heuristic data summarization algorithms based on different structured histograms. Last but not the least, we have conducted extensive experimental evaluations on the performance of our proposed solutions.