Skyline Predilection Query Based on Huge and Partial Dataset

S. Jagan[1], Dr. S. P. Rajagopalan[2]

[1]. Assistant Professor, Department of Computer Science and Engineering, Dr.APJ Abdul Kalam Centre for Research, Adhi College of Engineering and Technology, Kanchipuram-631 605, Tamil Nadu, INDIA

[2]. Professor, Department of Computer Science and Engineering, GKM College of Engineering and Technology, Chennai-600 063, Tamil Nadu, INDIA

Abstract

Personalized recommendation and the processing of real-time data illustrate the processing of huge data which in the field of Internet-of-Things (IoT) received a great extent of attention in recent literature. The partialness of huge data in the IoT is common. Obtaining personalized information from the partial data set is still puzzled by searching efficient and accurate methods at present. Skyline query is a widely used data processing method, particularly in the field of multi-objective decision analysis and data revelation. To eliminate the negative effects on huge data processing in IoT, an enhanced skyline predilection query strategy based on huge and the partial data set is proposed in this paper. This strategy simply separates divides huge and partial data set into two parts according to dimension importance and executes skyline query respectively. The strategy mainly resolves the problem of extracting personalized information from huge and partial data set and improves the efficiency of skyline query on huge and partial data set. Firstly, this paper presents a skyline predilection query strategy based on strict clustering and implements it on dimensions that have higher significance. Secondly, a skyline predilection query strategy based on loose clustering is implemented on dimensions that have lower significance. Finally, integrating local skyline query results, this paper calculates large-scale skyline query results by using information entropy theory. The efficiency and usefulness of SPQ algorithm have been evaluated in terms of response time and result set size during the comparative experiments with ISkyline algorithm and SIDS algorithm. A large number of simulation results show that the efficiency of SPQ algorithm is advanced than that of other common methods.

INTRODUCTION

The Internet of Things (IoT) and Cyber-Physical System(CPS) have made it possible for devices around the World to acquire information and store it, in order to be able to use it at a later stage [2], [3]. Currently, data is obtained by sensors and monitors in the field of IoT [4-6]. Due to the sensors and monitors failure, error and the existence of restrictions in acquiring actual data, misunderstanding data and missing values in datasets [7], the partialness of huge data is commonly observed in the field of Internet-of-Things [8], [9]. The dataset includes tuples have missing values in some of their dimensions, known as the partial data set. With the development and popularization of the Internet-of-Things, personalized recommendation aims to meet the needs of users becomes the hotspot of data processing in the field of Internet-of-Things. For example, according to the data from the smart bracelet, smart watches, and other wearable devices, different manufacturers can recommend their products for different users based on manufacturers predilection s. The problem is that how we obtain the suitable information, which meets the users' needs, in partial and huge dataset? Skyline query [10] is a typical multi-objective optimization method to solve this problem, which plays a key role in decision-making, market analysis, environmental monitoring, data mining, database visualization and econometrics [11]. Therefore, the skyline predilection query processing is a new angle and a cut-in point for solving the problem of the huge partial data from Internet-of-Things and Cyber-Physical System. In the past, the partial data set was cleaned, repaired or processed by any pre-treatment (e.g., see [12], [13], [14]) before skyline query [15], [16]. However, pre-treatment consumes too much system resources and resulted in enormous errors in the repaired data, which leads to inaccurate outcomes. Further, for some timeliness problems, such as processing the data in influenza period, pre-processing these huge and timeliness data may lead to slow response and data invalid. In this paper, the pre-treatment stage of the traditional method is abandoned, and a skyline predilection query strategy is proposed, which divided data set into two parts according to dimension importance: Skyline Predilection Query Based on Huge and Partial Dataset (SPQ). The execution efficiency of skyline query on the huge and partial data set has been significantly enhanced and personalized data that satisfies users predilection can also be obtained. The major contributions of this paper are as follows:

1. Our strategy gets rid of the pre-treatment stage that in general skyline query on partial data set, which reduces the response time of system.
2. We divide partial dataset according to the importance degree of attributes. Skyline query based on strict clustering on the projection of attributes have higher importance degree can ensure accuracy and skyline query based on loose clustering on the remaining of dataset can improve efficiency.
3. We define tuple encoding, strict cluster encoding and inclusion relation as the basis of SPQ algorithm.
4. We introduce two new algorithms, namely, SAVO algorithm and skyline predilection query based on dominance degree which are executed on two projections of partial dataset respectively.
5. We introduce a result selection strategy based on information entropy computation as a novel method designed especially for resolving the problem that the intersection of SSRS and RSRS is empty.
6. We give experimental evidence that the SPQ algorithm is efficient and individual, especially for huge and high-dimensional dataset.

The remainder of this paper proceeds as follows: Section II provides an overview of the related work in the literature. Encoding and clustering strategies are illustrated in Section III. In Section IV, two skyline predilection query strategies and a selection strategy based on information entropy are described. This is followed by simulation results and performance evaluation of SPQ presented in Section V. Finally, the paper is concluded in Section VI.