ABSTRACT

Navigating through multidimensional data cubes is a nontrivial task. Although On-Line Analytical Processing (OLAP) provides the capability to view multidimensional data through rollup, drill-down, and slicing-dicing, it offers minimal guidance to end users in the actual knowledge discovery process. In this article, we address this knowledge discovery problem by identifying novel and useful patterns concealed in multidimensional data that are used for effective exploration of data cubes. We present an algorithm for the Discovery of Sk-NAvigation Rules (DISNAR), which discovers the hidden interesting patterns in the form of Sk-navigation rules using a test of skewness on the pairs of the current and its candidate drill-down lattice nodes. The rules then are used to enhance navigational capabilities, as illustrated by our rule-driven system. Extensive experimental analysis shows that the DISNAR algorithm discovers the interesting patterns with a high recall and precision with small execution time and low space overhead.

Keywords: cube navigation; data cube lattice; OLAP; navigation rules; skewness

INTRODUCTION

With the ever-increasing volume of data collected and archived by organizations, it has become increasingly critical in order to navigate efficiently and effectively through large, multidimensional databases. Dimensional modeling techniques offer modeling paradigms in order to capture measures along multiple dimensions, and On-Line Analytical Processing (OLAP) tools provide various operations such as rollup, drill-down, and slicing-dicing in order to select target datasets and to view them from different angles. However, it is still a daunting task for end users to detect manually the hidden patterns in the voluminous and complex lattice of multidimensional databases. The manual data analysis during cube exploration becomes a bottleneck in the knowledge-discovery process.
In this article, we address this problem by proposing a knowledge-discovery technique in order to identify novel and useful patterns in multidimensional data that are later presented to end users in an apprehensible form. Specifically, we propose Discovery of Sk-NAvigation Rules (DISNAR), a novel skewness-based algorithm, in order to detect hidden surprises in data cubes, and then we use these surprises to provide a method for cube navigation. In this context, a surprise reveals how anomalous a set of transactions is when compared with another set of closely related transactions in the fact table. The anomalous transactions could be defined either by few outliers in the datasets influencing the aggregated datasets or by a group of transactions showing substantial difference in facts, such as profit or cost, from the remaining transactions. Because the notion of a surprise is an intuitive one, different users may have different impressions on what constitutes a surprise. Our rule-driven system allows users to control the knowledge-discovery process by letting them set the baseline for surprises by simply adjusting the level of significance.

Our work addresses two open research issues. The first one revolves around the inadequate level of decision support provided by most OLAP systems and is limited to aggregated data, which may not be sufficient for all users. Users often need to form concepts related to surprises in the data. The proposed approach using skewness aids in forming these concepts of surprises. The second one deals with the difficulty of navigating through data cube lattices. Currently, a user must have a fairly good understanding of the multidimensional model and a good intuition of what might be discovered in order to navigate through the vast magnitude of combinatorially explosive datasets involving high dimensionality and high variability in a data cube lattice. Without such knowledge, exploring these huge datasets is constrained by minimal system guidance and often misled by aggregated views. For instance, in a hypothetical profit scenario, a user may view that for years 1991 through 1998 the total profit values are very close to each other and may conclude that there is insignificant difference among the years. However, the data may contain some very high as well as very low profit segments when drilled down to lower hierarchical levels defined by quarters, months, and weeks, even though the aggregated values reveal no annual profit differences. This scenario describes a critical issue in cube exploration in which aggregated views with no definite guidance often present incorrect roadmaps to users. We approach this issue by guiding users based on surprises in the dataset. Thus, our methodology provides proper guidance through the discovery of surprises and, at the same time, lets users drive the knowledge exploration process.

This article is fundamentally different from related work in the sense that it identifies surprises by examining the data at the lowest level of granularity at every cuboid in a cube lattice, as opposed to the common practice of using aggregated data. Aggregation often hides the characteristics of the detailed data; an extremely high value and an extremely low value can be aggregated to a moderate value, hiding both extreme values. Using the smallest level of granularity, the problem of hiding a surprise as a side effect of aggregation is avoided. Our key contributions are as follows:

1. We discover hidden surprises with a high recall and precision in multidimensional cubes by using the statistical property of skewness on lattice nodes.
2. We enable users to examine different sets of surprises according to their needs. In our rule-driven system, a user can navigate to the same dataset multiple times by adjusting the critical level of significance of skewness. The practical implication of this adjustment is that some obvious surprises already may be known to the user, and therefore, he or she might search for more fine-grained surprises in different iterations.
3. We detect the surprises for multiple measures simultaneously at lattice nodes. For example, in most real-life settings, multiple measures such as revenue and
Literature Review in Computational Linguistics Issues in the Developing Field of Consumer Informatics: Finding the Right Information for Consumer’s Health Information Need
www.igi-global.com/chapter/literature-review-computational-linguistics-issues/21756?camid=4v1a

An Intelligent Data Mining System Through Integration Of Electromagnetism-Like Mechanism And Fuzzy Neural Network
www.igi-global.com/chapter/intelligent-data-mining-system-through/42356?camid=4v1a