Vertical Data Mining

William Perrizo
North Dakota State University, USA

Qiang Ding
Concordia College, USA

Qin Ding
Pennsylvania State University, USA

Taufik Abidin
North Dakota State University, USA

INTRODUCTION

The volume of data keeps increasing. There are many data sets that have become extremely large. It is of importance and a challenge to develop scalable methodologies that can be used to perform efficient and effective data mining on large data sets. Vertical data mining strategy aims at addressing the scalability issues by organizing data in vertical layouts and conducting logical operations on vertical partitioned data instead of scanning the entire database horizontally.

BACKGROUND

The traditional horizontal database structure (files of horizontally structured records) and traditional scan-based data processing approaches (scanning files of horizontal records) are known to be inadequate for knowledge discovery in very large data repositories due to the problem of scalability. For this reason, much effort has been put on sub-sampling and indexing as ways to address and solve the problem of scalability. However, sub-sampling requires that the sub-sampler know enough about the large dataset in the first place in order to sub-sample “representatively.” That is, sub-sampling requires considerable knowledge about the data, which, for many large datasets, may be inadequate or non-existent. Index files are vertical structures. That is, they are vertical access paths to sets of horizontal records. Indexing files of horizontal data records does address the scalability problem in many cases, but it does so at the cost of creating and maintaining the index files separate from the data files themselves.

A new way to organize data is to organize them vertically, instead of horizontally. Data miners are typically interested in collective properties or predictions that can be expressed very briefly (e.g., a yes/no answer). Therefore, the result of a data mining query can be represented by a bitmap vector. This important property makes it possible to do data mining directly on vertical data structures.

MAIN THRUST

Vertical data structures, vertical mining approaches and multi-relational vertical mining will be explored in detail to show how vertical data mining works.

Vertical Data Structures

The concept of vertical partitioning has been studied within the context of both centralized and distributed database systems for a long time, yet much remains to be done (Winslett, 2002). There are great advantages of using vertical partitioning: for example, it makes hardware caching work really well, it makes compression easy to do, and it may greatly increase the effectiveness of the I/O device since only participating fields are retrieved each time. The vertical decomposition of a relation also permits a number of transactions to be executed concurrently. Copeland & Khoshafian (1985) presented an attribute-level Decomposition Storage Model called DSM, similar to the Attribute Transposed File model (ATF) (Batory, 1979), which stores each column of a relational table into a separate table. DSM was shown to perform well. It utilizes surrogate keys to map individual attributes together, hence requiring a surrogate key to be associated with each attribute of each record in the database. Attribute-level vertical decomposition is also used in Remotely Sensed Imagery (e.g., Landsat Thematic Mapper...
Related Content

Interactive Quality-Oriented Data Warehouse Development
www.igi-global.com/chapter/interactive-quality-oriented-data-warehouse/28162?camid=4v1a

The VLEG Based Production and Maintenance Process for Web-based Learning Applications
Jorg Schellhase and Udo Winand (2002). Data Warehousing and Web Engineering (pp. 266-274).
www.igi-global.com/chapter/vleg-based-production-maintenance-process/7874?camid=4v1a

Positive Unlabelled Learning for Document Classification
www.igi-global.com/chapter/positive-unlabelled-learning-document-classification/11026?camid=4v1a

Employing Neural Networks in Data Mining
www.igi-global.com/chapter/employing-neural-networks-data-mining/10636?camid=4v1a