CubeExplorer: Online Exploration of Data Cubes

Jiawei Han[†‡] Jianyong Wang[†§] Guozhu Dong[¶] Jian Pei[‡] Ke Wang[‡]

[†] University of Illinois at Urbana-Champaign, U.S.A.
[‡] Simon Fraser University, Canada
[§] Peking University, China
[¶] Wright State University, USA

ABSTRACT

Data cube enables fast online analysis of large data repositories, which is attractive in many applications. Although there are several kinds of available cube-based OLAP products, users may still encounter challenges on effectiveness and efficiency in the exploration of large data cubes due to the huge *computation space* as well as the huge *observation space* in a data cube. CubeExplorer is an integrated environment for online exploration of data cubes. It integrates our newly developed techniques on *iceberg cube computation* [2], *cube-based feature extraction*, and *gradient analysis* [1], and makes cube exploration effective and efficient. In this demo, we will show the features of CubeExplorer, especially its power and flexibility at exploring and mining of large databases.

1. INTRODUCTION

As an integrated cube computation and cube exploration environment, CubeExplorer has the following distinct features which will be shown in our demonstration:

1. **Efficient computation of data cubes and iceberg cubes.** A concise data structure, called *H-tree* [2], has been developed for efficient computation of data cubes and iceberg cubes. Such a data structure and its associated cubing method (called *H-cubing*) are highly efficient for on-line computation of whole or part of data cubes. The effectiveness and efficiency of H-cubing at cube computation will be demonstrated.
2. **Efficient computation of iceberg cubes with complex measures.** Certain popular complex measures, such as `AVG()`, pose challenges to efficient iceberg cube computation. Based on our study in [2], such measures can be computed efficiently by exploring the corresponding weaker but anti-monotonic conditions to test and prune search space. The effectiveness and scalability of this approach at computing complex measures will be demonstrated.
3. **Mining multi-dimensional constrained gradients in data cubes.** With user-specified constraints, including probes, significant constraints and gradient constraints, multi-dimensional cube gradient analysis can be performed to identify interesting gradients in data cubes [1]. With this technique, users can explore online data cubes with gradient queries which are expressive, capable of capturing trends in data, and answering "what-if" questions. Besides computing complete gradients satisfying certain constraints, top-k interesting gradient cells can also be mined from data cubes efficiently. Mining top-k interesting gradient cells will relieve user's burden of specifying subtle constraint thresholds but return only high-quality answers. Both multi-dimensional gradient analysis and top-k gradient analysis will be presented in the demo.
4. **Predictive analysis by data cube exploration.** With the capability of top-k cube gradient analysis, one can obtain the general distribution of interesting gradients in a data cube. *How can we perform further analysis on such distribution and predict overall high quality gradients?* In this demo, we will show our method on predictive analysis of cube gradients, which leads to high quality prediction based on the results of gradient analysis. Furthermore, such analysis can be accompanied with drilling, rolling, dicing, and mutating, so that predictive analysis can be explored in different facets of a data cube, with a variety of user-defined measures.

Besides exploration of some typical data cubes, we will bring one real business database to show how CubeExplorer can be used for effective cube exploration and mining. Such a real database comes from an industry, with minor preprocessing that removes some subtle identities. This real data set will show the usefulness of CubeExplorer in applications.

* The work was supported in part by NSERC and NCE of Canada and the Univ. of Illinois at Urbana-Champaign.

ACM SIGMOD '2002 June 4-6, Madison, Wisconsin, USA

2. REFERENCES

[1] G. Dong, J. Han, J. Lam, J. Pei, and K. Wang. Mining multi-dimensional constrained gradients in data cubes. In *VLDB'01*, pp. 321–330, Rome, Italy, Sept. 2001.

[2] J. Han, J. Pei, G. Dong, and K. Wang. Efficient computation of iceberg cubes with complex measures. In *SIGMOD'01*, pp. 1–12, Santa Barbara, CA, May 2001.