Interactive Exploration and Visualization of OLAP Cubes

Carlos Ordonez
University of Houston
Houston, TX 77204, USA

Zhibo Chen
University of Houston
Houston, TX 77204, USA

Javier García-García
UNAM/IPN
Mexico City, Mexico

ABSTRACT

An OLAP cube is typically explored with multiple aggregations selecting different subsets of cube dimensions to analyze trends or to discover unexpected results. Unfortunately, such analytic process is generally manual and fails to statistically explain results. In this work, we propose to combine dimension lattice traversal and parametric statistical tests to identify significant metric differences between cube cells. We present a 2D interactive visualization of the OLAP cube based on a checkerboard that enables isolating and interpreting significant measure differences between two similar cuboids, which differ in one dimension and have the same values on the remaining dimensions. Cube exploration and visualization is performed by automatically generated SQL queries. An experimental evaluation with a medical data set presents statistically significant results and interactive visualizations, which link risk factors and degree of disease.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.3.6 [**Computer Graphics**]: Methodology and Techniques—*Interaction techniques*

General Terms

Algorithms, Human Factors

1. INTRODUCTION

On-Line Analytical Processing (OLAP) [2, 3, 7] is an important set of techniques for exploratory database analysis. In OLAP a large data set is analyzed with multiple aggregations to find interesting results. Such aggregations, computed based on multiple dimension combinations, resemble a multidimensional cube whose mathematical structure is represented by a lattice. In general, cube computations return simple descriptive statistics such as sums, row counts and averages. In this work, we show parametric statistical tests can help analyze the cube with high statistical reliability. On the other hand, we study how to visualize interesting results discovered from the cube.

From a data mining perspective, statistical tests [14] offer important advantages over statistical models. There are weak assumptions on the probabilistic distribution behind attributes from the data set. For the statistical test we used, a numeric attribute is assumed to have a Gaussian distribution. Statistical tests can produce statistically reliable results with large or small data sets, whereas most data mining models generally require large data sets to find significant results. Statistical tests are based on simple equations that can be efficiently evaluated with SQL queries because they generally do not require vector or matrix manipulation. By themselves, statistical tests generally require multiple trial and error runs before significant findings are produced. Moreover, each such run requires users to vary multiple parameters or select different dimension subsets from the data sets. Based on such issues, we developed algorithms that automate this process of exploring and analyzing a data set by combining OLAP cubes with parametric statistical tests. By enhancing standard exploratory OLAP techniques with statistical tests, we are able to prove that our findings are indeed significant, as opposed to obtaining findings from simple number comparisons. Our algorithms automatically analyze all cuboids from a multidimensional cube while applying statistical tests to discover significant differences in cube measure attributes. While our current application are medical databases, the algorithm enables the analysis of any OLAP database in which the goal is to find specific sets of dimensions that significantly change measure values. This problem is computationally challenging because the cube dimension lattice, which forms the foundation on which the statistical tests are based, represents a combinatorial search space. Due to this, we also introduce several algorithmic and systems optimizations that work towards improving the performance of this exhaustive comparison process.

The article is organized as follows. Basic definitions for OLAP databases and statistical tests, as well as an example are introduced in Section 2. The process by which we apply statistical tests on all the cuboids of a dimensional cube using OLAP is explained in Section 3. Section 4 explains our research on the visual analysis of the cube. Section 5 discusses related work. Section 6 presents conclusions and directions for future research.

2. DEFINITIONS

We now provide basic OLAP cube definitions. The input data is F, a fact table containing n records having d cube dimensions [7], $D = \{D_1, \ldots, D_d\}$ and a set of e measure [7] attributes $A = \{A_1, A_2, \ldots, A_e\}$. The mathematical structure representing all subsets of dimensions and their containment with each other is called the dimension lattice [7]. In

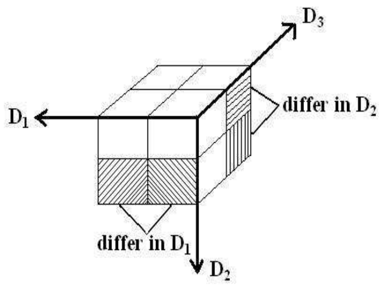


Figure 1: Discovering pairs of subsets with significant measure differences.

OLAP cube processing, the fundamental idea is to compute aggregations (sum(), count()) on measures A_i by subsets of dimensions (i.e. cuboids or cuboids) G s.t. $G \subseteq D$, effectively performing aggregations at varying granularity levels. The set of all potential aggregations at a certain level in the lattice (i.e. with a fixed number of dimensions) is called a cuboid and one specific aggregation group is called a cell.

In our algorithms, aggregations are used to derive univariate statistics such as μ, σ, which in turn are the basic input elements in the equations of a parametric statistical test. Section 3 explains in more detail cube exploration, the parametric statistical test and how they are combined.

Example

Figure 1 presents a simple example of a cube having three dimensions D_1, D_2, D_3 and one measure A_1. Each face of the cube represents a 2-dimensional cuboid. In this example, there are two sets of cell pairs within one cuboid that differ in exactly one dimension. The difference in fill pattern indicates there is a significant difference on a specific measure attribute A_1.

3. INTEGRATING OLAP CUBES AND STATISTICAL TESTS

3.1 Statistical Tests

Prior work [4, 6, 5] generally focused on looking for interesting or unusual patterns on single cells of the cuboids. As such, these works perform simple comparisons between the individual cells or between cells in different aggregations to obtain results. In contrast, we use a parametric statistical test to compare the population means [14] of pairs of groups. Our approach exhibits the following advantages: Two large groups of any size can be compared including groups with very different number of elements (e.g. a large and a small group). The means comparison test takes into account data variance, which measures overlap between the corresponding pair of populations. In the case of OLAP, dimensions can be used to focus on highly similar groups, differing in a few dimensions. It represents a natural extension of OLAP computations since it relies on distributive aggregations [5]. Cube measures are assumed to have a normal distribution, which is a reasonable assumption in most cases.

Let us know describe the means comparison parametric test in more formal terms. This statistical test compares the means, μ_1 and μ_2, from two similar but independent populations of sizes N_1 and N_2, respectively. We use a null hypothesis H_0 of $\mu_1 = \mu_2$ with the goal of finding pairs of populations in which H_0 can be rejected. In order to reject H_0, we aim to reach a high reliability (confidence) value $1 - p$, where p often calls in the following thresholds ($p \leq \{0.01, 0.05, 0.10\}$. When H_0 can be rejected with high confidence, we are able to accept the alternative hypothesis, H_1, which asserts the complementary comparison $\mu_1 \neq \mu_2$. Our parametric statistical test uses a two-tailed test which allows finding a significant difference on both tails of the Gaussian distribution. In order to determine the lowest possible p, we first determine the random variable z with a standard probabilistic distribution $N(0,1)$ based on Equation 1:

$$z = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}}, \qquad (1)$$

where μ_i, σ_i correspond to the estimated mean and standard deviation for groups (populations) $1, 2$, respectively. When N is large for both groups, it is only necessary to compare the z value with $z_{p/2}$ in the cumulative probability table for $N(0,1)$. When either group or both groups are small, we need to compute an additional value, the degrees of freedom: $df = N_1 + N_2 - 2$, which together with z, is used to lookup the p value on the t-student distribution table.

Performing multiple statistical tests on cuboids

We have two main objectives for the application of parametric statistical test: (1) Discovering significant differences between two groups in a cuboid on at least one measure. We should mention that finding two or more significant measure differences based on the same dimension combination is desirable, but unusual. (2) When there exists a significant difference we isolate those groups that differ in one dimension, which can explain a cause-effect relationship. Even though dimensions are assumed independent the aggregation automatically groups records with correlated dimensions together. Therefore, if a high correlation exists in binary dimensions it will be automatically considered.

With respect to goal (1), when applying a statistical test a significant difference can only be supported by a small p-value which takes into account both the means and the standard deviation of the distributions. The smaller the p-value the more likely the difference in the cube measure value between both groups is significant. It is expected that many measure differences will not be significant, making the search problem on the dimension lattice expensive. With respect to goal (2), the algorithm aims to discover significant differences in highly similar cube cells because that helps point out which specific dimension “triggers” a significant change on the cuboid measure. In other words, finding a significant measure difference, between two highly dissimilar groups, makes a cause-effect explanation difficult, since such difference may be caused by the interaction of two or more cube dimensions. However, those less significant measure differences can be stored on additional tiers.

3.2 Exploration and Visualization Algorithm

We introduce an algorithm that integrates cube exploration, statistical tests and visualization. This algorithm extends our previous algorithm [11] with visualization and interactive exploration. Our algorithm has the following goals: (1) exploring all cuboids from F (when $d \leq 10$). Otherwise, exploring all cuboids based on k dimensions manually selected by the user s.t. $k < d$; (2) performing the statistical test for every cube pair; (3) selecting significant pairs differing in $\delta = 1$ cube dimensions; (4) interactive visual exploration of the cube, together with statistically significant results; (5) efficient visualization of associated image data per cube cell.

Our algorithm basically computes the entire cube, exploring the entire dimension lattice and then applies statistical tests for every pair. The algorithm assumes a low d or alternatively low k, binary dimensions, which is common in medical databases. Our tool applies a top-down approach exploring all cuboids from a cube, working level-wise. Further details can be found in [11].

4. VISUAL ANALYSIS OF THE CUBE

In this section, we explain interactive exploration and visualization of a medical set with our algorithms. First, we explain our computer and DBMS setup and the input medical data set used as a fact table. Second, we explain lattice exploration guided by visualization of cuboids at different dimensionalities. Third, we explain how we visualize significant cuboid cell pairs as well as associated attributes, including image data.

4.1 Computer Setup and Data Set

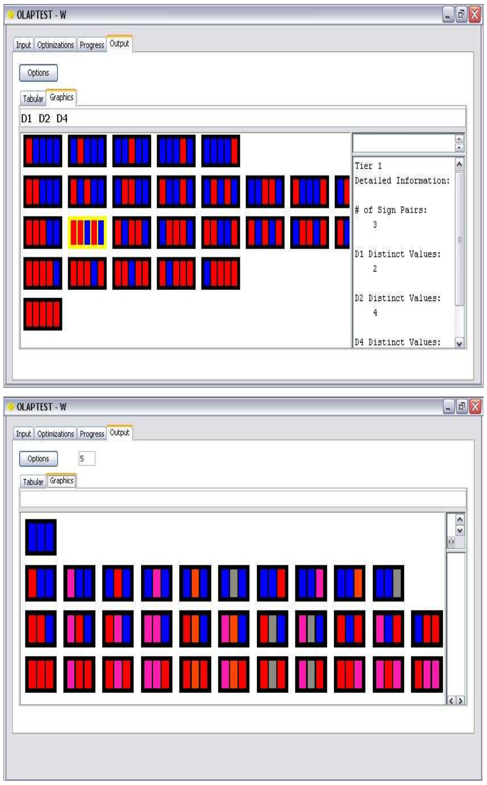
Our system was developed in Java that automatically generates SQL queries. JDBC is used to connect to the DBMS. We used the SQL Server DBMS, running on a computer with 3.2GHz CPU, 4GB of RAM, and 1TB disk.

We performed experiments on a real data set coming from the medical domain. Our medical data set contains profiles of $n = 655$ patients and has 25 attributes containing categorical, numeric and image data. There were medical measurements such as weight, heart rate, blood pressure and pre-existence of related diseases. Finally, the data set contains the degree of artery narrowing (stenosis) for the four heart arteries. All numeric attributes were converted to binary dimensions. There were $d = 12$ binary dimensions (e.g. gender, hypertension Y/N), $e = 4$ measures (artery disease measurement) $f = 9$ image attributes representing a standardized image of the heart.

We now explain parameter settings. For our medical data sets our goal was to explore the entire dimension lattice. Therefore, we used all d dimensions. The settings for parameters were as follows. $p = 0.01$, $\delta = 1$, which can be interpreted as follows. We want to find significant measure differences, with 99% confidence, on all group pairs differing in one dimension. A group pair in the cube can have from 1 to d dimensions, out of which one will be different. It is possible, but unlikely, that a group pair has significant differences in two or more measures.

4.2 Application in Medicine

Table 1 shows actual significant findings on the medical data set, introduced above. Each row represents the comparison between two patient groups, differing in one dimension indicated by “0/1”. Cube dimensions are risk factors for heart disease including family history of heart disease, diabetes, gender, high cholesterol and high blood pressure. For a given group, each matching dimension will be ”0” indicating absence of a risk factor, ”1” indicating presence of a risk factor, or “All” when such dimension was ignored in the aggregation. When p is small it indicates two highly similar groups of patients, differing in exactly one risk factor have a highly different degree of disease (artery narrowing).



D1 D2 D4

Tier 1
Detailed Information:

# of Sign Pairs:
3

D1 Distinct Values:
2

D2 Distinct Values:
4

D4 Distinct Values:

5

Figure 2: Visualization of cuboids.

4.3 Visualizing Cuboids

Our visualization architecture can be separated into two main objectives: visualization of cuboids and visualization of significant cell pairs.

In order to visualize the cuboids and allow for quick navigation between different cuboids, we created a two-tiered design, as shown in Figure 2. In this design, the first-tier represents a lattice-like display that shows the various combinations of the k dimensions through the use of blocks of items. Each item represents one specific dimension and can be colored red, to represent on, or blue, to represent off. A set of k items represents one possible combination of the chosen dimensions and is called a block. Since OLAP is an exhaustive process that traverses all possible combinations of dimensions, the total number of blocks is 2^k. We chose to arrange the blocks by the number of dimensions that are in the cuboid because it allows for easy navigation. In addition to just viewing the blocks, additional information, such as the total number of significant pairs, is also provided with mouse navigation. Since the k dimensions are inputted by the user prior to the execution of the algorithm, this first-tier does not require any retrieval of data from the database.

Once a block is selected, the cuboid that it represents will be displayed to show the second tier of our design. In this level, as shown in Figure 2, on the lower screenshot, there still exists block and items. However, in this case, each

Table 1: Medical data set: group pairs with significant measure differences.

D_1	D_2	D_3	D_4	D_5	N_1	N_2	A_1	A_2
FamHist	Diab	Gender	HighChol	highBP			LAD	RCA
0	All	All	1	0/1	35	23	$p > 0.1$	$p \in [0.01 - 0.05]$
0/1	All	All	1	0	35	26	$p > 0.1$	$p < 0.01$
All	0/1	All	All	All	47	157	$p < 0.01$	$p < 0.01$

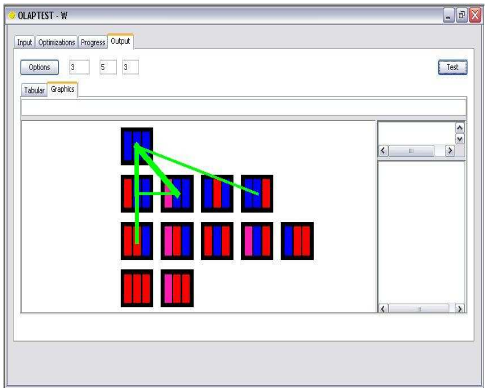


Figure 3: Visualization of Significant Cell Pairs.

block represents one specific combination of the selected dimensions in the cuboid, while each item represents a specific distinct value of the dimension. We use a different color to represent each distinct value for a certain dimension. As such, a red item in one dimension would represent a different value than a red item for another dimension. We provide a collapsible legend that provides information on the values that are represented by each color for each dimension. Note that this more detailed level not only contains the specific values for the selected cuboid, but also shows those blocks that form a significant pair with a block in the chosen cuboid. Once again, placing the mouse over a block will bring up additional information on the panel to the right. Selecting a specific block in this tier will bring up statistical data regarding this population, such as the mean and standard deviation. Additionally, a sample of actual records in the original data set that belong to this block can be viewed. In order to generate the visualization of the specific cuboid, we perform a single retrieval on the final result set with the application of appropriate predicates to reduce the size of the returned values. Since the final result table is indexed on all dimensions, we can quickly retrieve the required records. From this set, we can determine which distinct dimensions were involved in a significant cell pair.

4.4 Visualization of Significant Cell Pairs

The second portion of our visualization goal is to easily view and navigate the significant pairs returned by the algorithm. In our application, significant results are represented by a green line linking two blocks of values together, as can be seen in Figure 3. The thickness of the line represents the amount of significance between the two blocks. Should more than one measure be significant between any two blocks, then the thickness would also be altered to reflect this situation. As with the previous visualizations, placing the mouse over the line would provide additional information, such as p-value. Selecting a line would provide the user with detailed information regarding the overlap between the Gaussians of the two blocks, or populations, as seen in Figure 3. In this case, the panel to the right would display a graph showing two Gaussians, each representing one of the blocks. The user can thus graphically view the overlap present with the notion being that the smaller the amount of overlap, the greater the significance of the result. Extra information, such as mean and standard deviation, would also be provided on the right panel. In addition, we also provide a feature that allows for the selection of multiple significant lines. In this case, the right panel would display multiple sets of Gaussians to allow the user to visually discern the difference in overlap between different pairs. The creation of these connecting lines requires a single-pass on the final result set. We apply a filter on this result set to only return those significant pairs whose dimensions fit within the selected cuboid. Since the DBMS is optimized for fast retrieval, the time to obtain and generate this visualization is also very efficient. Once the data is obtained, they are stored in either main memory or, if the results are too large, are stored in a binary file for quick access. In this way, the Gaussian curves that are shown when a line is selected can be quickly displayed since the required information is already in RAM memory.

5. RELATED WORK

Cube exploration is a well researched topic. OLAP and a classification of aggregations originates in the seminal paper [5]. In [6] the authors put forward the plan of creating smaller, indexed summary tables from the original large input table to speed up aggregating executions. In [13] the authors explore tools to guide the user to interesting regions in order to highlight anomalous behavior while exploring large OLAP data cubes. This is done by identifying exceptions, that is, values in cells of a data cube that are significantly different from the value anticipated, based on a statistical model. In contrast, we propose to use statistical tests to do pair-wise comparison of neighboring cells in cuboids to discover significant metric differences between similar groups. We identify such differences giving statistical evidence about the validity of findings.

There has been research on visualizing OLAP cubes. Recent work can also be found on visually and interactively exploring data warehouses. The authors for [16] explore the requirements for analyzing a spatial database with an OLAP tool. This work shows the need to apply spatial data techniques, used in geographic information systems, for OLAP exploration, in which drill up/down, pivoting, and slicing and dicing provide a complementary perspective. In contrast, our work relies on statistical tests to explore OLAP cubes and can automatically detect significant metric differences between highly similar groups. Additional visualization work was completed in [9] where the mapping of the Cube Presentation Model, a display model for OLAP screens, involves visualization techniques from the Human-Computer Interaction field. In [8], the author presents a rigorous multidimensional visualization methodology for visualizing n-dimensional geometry and its applications to visual and automatic knowledge discovery. The application of visual knowledge discovery techniques is possible by transforming the problem of searching for multivariate relations among the variables into a two-dimensional pattern recognition problem. A framework for exploration of OLAP data with user-defined dynamic hierarchical visualizations is presented in [15]. Even though this study emphasizes the use of visualization tools to explore data warehouses, we propose a tool that not only gives the user visual aids to explore the data, but also to present the user with a novel method of highlighting interesting features of the cubes by means of statistical tests. Indexing is one method of increasing performance in the searching of images and the authors in [1] proposed a multilevel index structure that can efficiently handle queries on video data. Our work is related to applying data mining in medical data sets to improve heart disease diagnosis [12, 10].

6. CONCLUSIONS

We presented an innovative system that combines the exploratory power of OLAP cubes with the statistical reliability of statistical tests. Cube exploration is used to automatically analyze all subsets of dimensions, freeing the user from having to manually set parameters. Pairs of similar groups are isolated and compared with a statistical test to discover specific pairs that cause a significant difference in some measure value. Such differences represent statistically reliable results. The OLAP cube is depicted using a two-tier design that allows the user to quickly switch between cuboids. Significant results are visually shown using connecting lines and also provide additional information with one or more Gaussian graphs that graphically show the overlap between two record populations. We presented an application in the medical domain to improve heart disease diagnosis.

Cube visualization is a fertile research topic since it involves understanding a multidimensional data set. A 2D cube representation is easier to manipulate than a 3D display, but we would like to compare its strengths and weaknesses with a 3D visualization. We need to study mathematical relationships between the dimensions lattice and cube visual representations. The visualization of cubes requires further study when confronted with a cube having a large number of dimensions. Finally, we want to integrate other statistical tests or statistical models with OLAP cubes.

Acknowledgments

This research work was partially supported by grant NSF IIS 0914861 from the National Science Foundation. Javier García-García affiliation institutions are Universidad Nacional Autónoma de México (UNAM), CU campus and Instituto Politécnico Nacional (IPN), UPIICSA campus. We would like to thank the Emory University Hospital for providing the medical data set used in this work.

7. REFERENCES

[1] L. Chen, M. Ozsu, and V. Oria. Mindex: An efficient index structure for salient-object-based queries in video databases. *Multimedia Syst.*, 10(1):56–71, 2004.

[2] Z. Chen and C. Ordonez. Efficient OLAP with UDFs. In *Proc. ACM DOLAP Workshop*, pages 41–48, 2008.

[3] Z. Chen, C. Ordonez, and C. Garcia-Alvarado. Fast and dynamic OLAP exploration using UDFs. In *Proc. ACM SIGMOD Conference*, pages 1087–1090, 2009.

[4] L. Fu and J. Hammer. Cubist: a new algorithm for improving the performance of ad-hoc OLAP queries. In *Proc. ACM DOLAP Workshop*, 2000.

[5] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-total. In *ICDE Conference*, pages 152–159, 1996.

[6] H. Gupta, V. Harinarayan, A. Rajaraman, and J.D. Ullman. Index selection for OLAP. In *IEEE ICDE Conference*, 1997.

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 1st edition, 2001.

[8] A. Inselberg. Visualization and knowledge discovery for high dimensional data. In *UIDIS*, pages 5–24, 2001.

[9] A. S. Maniatis, P. Vassiliadis, S. Skiadopoulos, and Y. Vassiliou. Advanced visualization for OLAP. In *Proc. ACM DOLAP Workshop*, pages 9–16, New York, NY, USA, 2003. ACM Press.

[10] C. Ordonez. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine (TITB)*, 10(2):334–343, 2006.

[11] C. Ordonez and Z. Chen. Evaluating statistical tests on OLAP cubes to compare degree of disease. *IEEE Transactions on Information Technology in Biomedicine (TITB)*, 13(5):756–765, 2009.

[12] C. Ordonez, N. Ezquerra, and C.A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems (KAIS)*, 9(3):259–283, 2006.

[13] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *EDBT*, pages 168–182. Springer-Verlag, 1998.

[14] M. Triola. *Essentials of Statistics*. Addison Wesley, 2nd edition, 2005.

[15] S. Vinnik and F. Mansmann. From analysis to interactive exploration: Building visual hierarchies from OLAP cubes. In *EDBT*, pages 496–514, 2006.

[16] A. Voß, V. Hernandez, H. Voß, and S. Scheider. Interactive visual exploration of multidimensional data: Requirements for CommonGIS with OLAP. In *DEXA Workshops*, pages 883–887, 2004.