Evaluation of a MCA-Based Approach to Organize Data Cubes

Riadh Ben Messaoud
rbenmessaoud@eric.univ-lyon2.fr

Omar Boussaid
omar.boussaid@univ-lyon2.fr

Sabine Loudcher Rabaséda
sabine.loudcher@univ-lyon2.fr

Laboratory ERIC – University of Lyon 2
5 avenue Pierre Mendès–France,
69676 Bron Cedex, France

ABSTRACT

On Line Analysis Processing (OLAP) is a technology basically created to provide users with tools in order to explore and navigate into data cubes. Unfortunately, in huge and sparse data volumes, exploration becomes a tedious task and the simple user's intuition or experience does not always lead to efficient results. In this paper, we propose to exploit the results of the Multiple Correspondence Analysis (MCA) in order to enhance a data cube representation. Our approach address the issues of organizing data in an interesting way and detecting relevant facts. We also treat the problem of evaluating the quality of data representation in a multidimensional space. For this, we propose a new criterion to measure the relevance of data representations. This criterion is based on the concept of geometric neighborhood and similarity between cells of a data cube. The experimental results we led on real data samples have shown the interest and the efficiency of our approach.

Categories and Subject Descriptors

B.8.2 [**Hardware**]: Performance and reliability—*Performance Analysis and Design Aids*; E.1.1 [**Data**]: Data structures—*Arrays*; H.4.2 [**Information Systems**]: Information systems ApplicationsTypes of Systems[Decision support]

General Terms

Algorithms, Experimentation, Performance

Keywords

OLAP, Data cubes, Data representation, MCA, Test-values, Arrangement of attributes, Characteristic attributes, Homogeneity criterion

CIKM'05, 31^{st} October – 5^{th} November, 2005 Bremen, Germany

1. INTRODUCTION

On-Line Analytical Processing (OLAP) is a technology supported by most data warehousing systems [9, 12]. It provides a platform for analyzing data according to multiple dimensions and multiple hierarchical levels. Data are presented in multidimensional views, commonly called data cubes [3]. A data cube can be considered as a space representation composed by a set of cells. A cell is associated with one or more measures and identified by coordinates represented by one member from each dimension. Each cell in a cube represents a precise fact. For example, if dimensions are *products*, *stores* and *months*, the measure of a particular cell can be the *sales* of one *product* in a particular *store* on a given *month*. OLAP provides the user with tools to summarize, explore and navigate into data cubes in order to detect interesting and relevant information. However, exploring a data cube is not always an easy task to perform. Obviously, in large cubes containing sparse data, the whole analysis process becomes tedious and complex. In such a case, an intuitive exploration based on the user's experience does not quickly lead to efficient results. More generally, in the case of a data cube with more than three dimensions, a user is naturally faced to a hard task of navigation and exploration in order to detect relevant information. Current OLAP provides query-driven tools to browse data cubes, but does not deeply assist the user and help him/her to investigate interesting patterns.

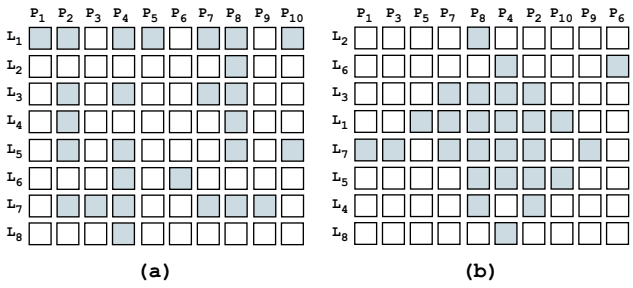


Figure 1: Example of different representations of a 2-dimensional data cube.

For example, consider the cube of Figure 1(a). This cube displays sales of products $(P_1, \ldots, P_{10})$ crossed by geographic locations of stores $(L_1, \ldots, L_8)$. In one hand, in the Fig-

ure 1(a), the way the cube is displayed does not provide an attractive representation that helps a user to easily interpret data. Full cells (gray cells) of are displayed randomly according to the lexical ordering of the members[1] in each dimension. On the other hand, Figure 1(b) contains the same information as Figure 1(a). However, it displays a data representation that is easier to analyze. Furthermore, the cube of Figure 1(b) emerges interesting facts by gathering full cells together. Such a representation is more comfortable to perform efficient analysis. Representation (b) of Figure 1 can be interactively constructed by the user from representation (a) via some classic OLAP operators. This suppose that the user intuitively knows how to arrange the attributes. Therefore, we propose to provide the user with an assistance to identify interesting facts and arrange them in a suitable representation. As shown in Figure 1, we propose an approach that allows the user to get relevant facts and displays them in an appropriate way that enhances the exploration process independently of the cube's size. Thus, we suggest to carry out a Multiple Correspondence Analysis [10] (MCA) on a data cube as a preprocessing step. Basically MCA is a powerful describing method even for huge volumes of data. It factors categorical variables and displays data in a factorial space constructed by orthogonal system of axis that provides relevant views of data. These elements motivate us to exploit the results of the MCA in order to better explore large data cubes by identifying and arranging its interesting facts. The firsts constructed factorial axis summarizes the maximum of information contained in the cube. We focus on relevant OLAP facts associated with characteristic attributes (variables) given by the factorial axis. These facts are interesting since they reflect relationships and concentrate a significant information. For a better visualization of these facts, we highlight them and arrange their attributes in the data space representation by using the *test-values* [15] (see subsection 5.1).

We also propose in this paper a novel quality representation criterion to evaluate relevance of multidimensional data representations (see section 6). This criterion is based on geometric neighborhood of data cube cells. It also take into account the similarity of measure of cells. The goal of this criterion is to provide a scalar quantification for the quality representation of a given data cube. It also allows to evaluate the performance of our approach by comparing the quality of initial data representation and the arranged one.

This paper is organized as follows. In section 2, we present some related work to our approach. We provide in section 3 the problem formalization and present the general context of this work. The section 5 introduces the *test-values* and details steps of our approach. We define in the next section our quality representation criterion. The section 7 presents a real world case study on a huge and sparse data cube. We propose experimental results in the section 8. Finally, we conclude and propose some future works.

2. RELATED WORK

Several works have already treated the issue of enhancing the space representation of data cubes. These works were undertaken following different motivations and adopted different ways to address the problem. While some are interested to technical optimization (storage space, queries response time, etc.), others have rather focused on OLAP aspects. Our present work fits into the second category. Recall that, in our case, we focus on assisting OLAP users in order to improve and help the analysis process on large and sparse data cubes. We use a factorial approach to highlight relevant facts and provide interesting data representations for the analysis.

In [21], Vitter *et al.* proposed to build compact data cubes by using approximation through wavelets. Another data structure, called `Quasi-Cube` [1], compresses data representation by materializing only sufficient parts of a data cube, the remaining parts are approximated by a linear regression. In [19] approximation is performed by estimating the density function of data. `Dwarf` [20] reduces the storage space of a cube by identifying and factoring redundant tuples in the fact's table. Wang *et al.* propose to factorize these redundancies by exploiting `BST` [22] (*Base Single Tuple*). Thus, a more condensed data cube (`MinCube`) was proposed. In [7], Feng *et al.* introduce `PrefixCube`, a data structure based on only one `BST`.

The `Quotient Cube` [13] method summarizes semantic contents of a data cube and partitions it into cells with identical values. In [14], `Quotient Cube` was involved and a novel data structure, `QC-Tree`, was proposed. `QC-Tree` is directly constructed from the base table in order to maintain it under updates. Feng *et al.* [8] identify correlation between attributes values and propose the `Range CUBE` method to compute and compress a data cube without loss of precision. Ross and Srivastava [18] propose `Partitioned-Cube`. This algorithm is based on partitioning the large relations into fragments. Operations over the whole cube are performed on each memory-sized fragment independently. In [16], huge high dimensional data are partitioned in disjoint small datasets. For each dataset, a local data cube is computed offline and used to compute queries in an online fashion.

Finally, in our approach we share already the same motivation of Choong *et al* [5, 4]. They also address the problem of high dimensionality of data cubes and try to enhance analysis processes by preparing the data set into appropriate representation so that the user can explore it in a more effective manner. The authors use an approach that combines association rules algorithm and a fuzzy subsets method. Their approach consists in identifying blocks of similar measures in the data cube. However, this approach does not take into account the problem of sparse data cubes.

3. PROBLEM FORMALIZATION

Let $\mathcal{C}$ denote a data cube. Note that, our approach can be applied directly on $\mathcal{C}$ or on a data view (a sub-cube) taken from an initial cube $\mathcal{C}$. It is up to the user to select dimensions, fix one hierarchical level per dimension and select measures in order to create a particular data view (s)he wishes to visualize. Thus, to enhance the data representation of the constructed view, the user can apply on it our proposed approach. In order to lighten the formalization, in the followings of the paper, we assume that a user has selected a data cube $\mathcal{C}$, with d dimensions $(D_1, \ldots, D_t, \ldots, D_d)$, m measures $(M_1, \ldots, M_q, \ldots, M_m)$ and n facts. We also assume that the user has fixed one hierarchical level with p_t categorical attributes per dimension. Let a_j^t the j^{th} attribute of the di-

[1]To avoid confusion in the followings, we adopt the term "attribute" to indicate a "member" or a "modality" of a dimension.

mension D_t and $p = \sum_{t=1}^{d} p_t$ the total number of attributes in $\mathcal{C}$. For each dimension D_t, we note $\{a_1^t, \ldots, a_j^t, \ldots, a_{p_t}^t\}$ the set of its attributes.

In a first step, the aim of our approach is to organize the space representation of a given data cube $\mathcal{C}$ by arranging the attributes of its dimensions. For each dimension D_t, our approach establishes a new arrangement of its attributes a_j^t in the data space (see subsection 5.2). This arrangement provides a data representation visually easier to interpret and displays multidimensional information in a more suitable way for analysis. In a second step, our approach detects from the resulted representation relevant facts expressing interesting relationships. To do that, we select from each dimension D_t a subset Φ_t of significant attributes, also called characteristic attributes (see subsection 5.3). The crossing of these particular attributes allows to identify relevant cells in the cube.

Our approach is based on the MCA [10, 15]. The MCA is a factorial method that displays categorical variables in a property space which maps their associations in two or more dimensions. From a table of n observations on p categorical variables, describing a p-dimensional cloud of individuals ($p < n$), the MCA provides orthogonal axis to describe the most variance of the whole data cloud. The fundamental idea is to reduce the dimensionality of the original data thanks to a reduced number of variables (factors) which are a combination of the original ones. The MCA is generally used as an exploratory approach to unearth empirical regularities of a dataset.

In our case, we assume the cube's facts as the individuals of the MCA, the cube's dimensions as its variables, and the attributes of a dimension as values of their corresponding variables. We apply the MCA on the n facts of the cube $\mathcal{C}$ and use its results to build *test-values* (see subsection 5.1) for the attributes a_j^t of the dimensions D_t. We exploit these *test-values* to arrange attributes and detect characteristic ones in their corresponding dimensions.

4. APPLYING THE MCA ON A DATA CUBE

Like all statistic methods, the MCA needs a tabular representation of data as input. Therefore, we can not apply it directly on a multidimensional representation. So, we should transform the data of $\mathcal{C}$ under a *complete disjunctive table*. For each dimension D_t, we generate a binary matrix Z_t with n lines and p_t columns. Lines represent the facts, and columns represent the dimension's attributes. The i^{th} line of Z_t contains $(p_t - 1)$ times the value 0 and one time the value 1 in the column that fits with the attribute taken by the fact i. The general term of Z_t is:

$$z_{ij}^t = \begin{array}{ll} 1 & \text{if the fact } i \text{ takes the attribute } a_j^t \\ 0 & \text{else} \end{array}$$

By merging the d matrices Z_t, we get the complete disjunctive table $Z = [Z_1, Z_2, \ldots, Z_t, \ldots, Z_d]$ having n lines and p columns. Z describes the d positions of the n facts of $\mathcal{C}$ through a binary coding. In the case of a large data cube, we naturally get very huge matrix Z. Recall that the MCA, like all factorial methods, is perfectly suitable for huge input dataset with high numbers of lines and columns.

Already having the complete disjunctive table Z, the MCA starts by constructing the *Burt* table $B = Z'Z$ (Z' is the transposed matrix of Z). B is a (p, p) symmetric matrix that contains all the category marginals on the main diagonal and all possible cross-tables of the d dimensions of $\mathcal{C}$ in the off-diagonal. Let consider X a (p, p) diagonal matrix that has the same diagonal elements of B and zeros otherwise. By diagonalizing the matrix $S = \frac{1}{d} Z'ZX^{-1}$, we obtain $(p - d)$ diagonal elements. These elements are called *eigenvalues* and noted λ_α. Each eigenvalue λ_α is associated to a directory vector u_α for a factorial axis F_α, where $Su_\alpha = \lambda_\alpha u_\alpha$. The Figure 2 summarizes the previous approach via the algorithm *CubeToMCA*. This algorithm creates a complete disjunctive table from an input cube $\mathcal{C}$, applies on the MCA and returns eigenvalues as output.

```
Algorithm CubeToMCA(C)
Input:
  C: data cube
Begin
  for (t = 1; t ≤ p; t++) do
    Z_t ← 0;
    for each attribute a_j^t in D_t do
      for each fact i in C do
        if (fact i takes a_j^t) then
          z_ij^t ← 1;
          Break for;
        end if
      end for
    end for
    Z ← merge(Z,Z_t);
  end for
  B ← ZZ';
  for (i = 1; i ≤ p; i++) do
    for (j = 1; j ≤ p; j++) do
      if (i ≠ j) then
        x_ij ← 0;
        else   x_ij ← b_ij;
      end if
    end for
  end for
  S ← (1/d) Z'ZX^{-1};
  S ← diagonalize(S);
  for (α = 1; α ≤ p − d; α++) do
    λ_α ← s_αα;
  end for
End
```

Figure 2: Algorithm CubeToMCA.

An eigenvalue represents the amount of inertia (variance) that reflects the relative importance of its axis. The first axis always explains the most inertia and has the largest eigenvalue. Usually, in a factorial analysis process, researchers keep only the first, two or three axis. Other researchers give complex mathematical criteria [2, 11, 17, 6] to determine the number of axis to keep. In [10], Benzecri suggests that this limit should be fixed by the user's capacity to give a meaningful interpretation to the axis he keeps. It's not because an axis has a relatively small eigenvalue that we should discard it. It can often help to make a fine point about the data. It's up to the user to choose the number k of axis (s)he wants to keep after checking the eigenvalues and the general meaning of the axis.

5. ORGANIZING DATA CUBES AND DETECTING RELEVANT FACTS

Usually in a factorial analysis, relative contributions of variables are used to give sense to the axis. A relative contribution shows the percent of inertia of a particular axis which is explained by an attribute. The largest relative contribution of a variable to an axis is, the more it gives sense

of this axis. In our approach, we interpret a factorial axis by characteristic attributes detected through the use of the *test-values* proposed by Lebart *et al.* in [15]. We present in the following subsection the theorical principle of test-values applied to the context of our approach.

5.1 The test-values

Let $I(a_j^t)$ denotes the set of facts having a_j^t as attribute in the dimension D_t. We also note $n_j^t = \text{Card}(I(a_j^t)) = \sum_{i=1}^{n} z_{ij}^t$ the number of elements in $I(a_j^t)$. It corresponds to the number of facts in $\mathcal{C}$ having a_j^t as attribute (weight of a_j^t in the cube). We consider $\varphi_{\alpha j}^t = \frac{1}{n_j^t \sqrt{\lambda_\alpha}} \sum_{i \in I(a_j^t)} \psi_{\alpha i}$ the coordinate of a_j^t on the factorial axis F_α, where $\psi_{\alpha i}$ is the coordinate of the facts i on F_α. Suppose that, under a null hypothesis H_0, the n_j^t facts are selected randomly in the set of the n facts, the mean of their coordinates in F_α can be represented by a random variable $Y_{\alpha j}^t = \frac{1}{n_j^t} \sum_{i \in I(a_j^t)} \psi_{\alpha i}$, where $\text{E}(Y_{\alpha j}^t) = 0$ and $\text{VAR}_{H_0}(Y_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{\lambda_\alpha}{n_j^t}$. Remark that $\varphi_{\alpha j}^t = \frac{1}{\sqrt{\lambda_\alpha}} Y_{\alpha j}^t$. Thus, $\text{E}(\varphi_{\alpha j}^t) = 0$, and $\text{VAR}_{H_0}(\varphi_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{1}{n_j^t}$. The test-value of the attribute a_j^t is:

$$V_{\alpha j}^t = \sqrt{n_j^t \frac{n-1}{n - n_j^t}} \varphi_{\alpha j}^t \tag{1}$$

$V_{\alpha j}^t$ measures the number of standard deviation between the attribute a_j^t, i.e. the gravity center of its n_j^t facts, and the center of factorial axis F_α. The position of an attribute is interesting for a given axis F_α if its cloud of facts is located in a narrow zone in the direction α. This zone should also be as far as possible from the center of the axis. The test-value is a criterion that quickly provides an appreciation if an attribute has a *significant* position on a given factorial axis or not.

5.2 Arrangement of attributes

In a classic OLAP representation of data cubes, attributes of dimensions are usually organized according to a lexical order such as alphabetic order for *geographic* dimensions or chronological order for *times* dimensions. In our approach, we propose to exploit the test-values of attributes in order to organize differently the data cube's facts. The new organization will display a relevant data representation easier to analyze and to interpret especially in the case of large and sparse cubes. For each dimension of a data cube, we sort its attributes according to the increasing order of their test-values. Actually, a test-value indicates the position of an attribute on a given axis. The relative geometric position of an attribute is more significant to a factorial axis when these axis are important (have the greatest eigenvalues). For this, we propose to sort attributes according to the k first axis selected by the user. We sort the p_t test-values $V_{\alpha j}^t$ of the attributes a_j^t on the axis F_α. This will provide a new order of indices j. According to this order, we arrange attributes a_j^t in the dimension D_t.

In general, we assume that all attributes of a dimension D_t are geometrically ordered in the data cube space representation according to the order of indices j_t. i.e, the attribute $a_{j_t-1}^t$ precedes $a_{j_t}^t$ and $a_{j_t}^t$ precedes $a_{j_t+1}^t$ (see the example of Figure 3). Indices j_t are ordered according to the arrangement of the attributes in the space representation of the dimension D_t. Let us take an example of a dimension D_t with four attributes $\{a_1^t, a_2^t, a_3^t, a_4^t\}$. In the Table 1(a) attributes are arranged according to the initial order in the space representation. Therefore, the Table 1(b) displays a new arrangement of attributes by sorting its test-values on F_1 then by test-values on F_2.

Attributes	a_1^t	a_2^t	a_3^t	a_4^t
Test-values on F_1	$V_{11}^t = 5$	$V_{12}^t = 11$	$V_{13}^t = 7$	$V_{14}^t = 5$
Test-values on F_2	$V_{21}^t = 12$	$V_{22}^t = 3$	$V_{23}^t = 4$	$V_{24}^t = 7$

(a)

Attributes	a_4^t	a_1^t	a_3^t	a_2^t
Test-values on F_1	$V_{14}^t = 5$	$V_{11}^t = 5$	$V_{13}^t = 7$	$V_{12}^t = 11$
Test-values on F_2	$V_{24}^t = 7$	$V_{21}^t = 12$	$V_{23}^t = 4$	$V_{22}^t = 3$

(b)

Table 1: Example of attributes (a) before and (b) after arrangement.

5.3 Characteristic attributes

In general, an attribute is considered significant for an axis if the absolute value of its test-value is higher than $\tau = 2$. This corresponds roughly to an error threshold of 5%. We note that, the lower error threshold is, the greater τ is. In our case, for one attribute, the test of the hypothesis H_0 can induce a possible error. This error will inevitably be increased when we perform the test p times for all the attributes of the cube. To minimize this accumulation of errors, we propose to fix for each test an error threshold of 1% which correspond to $\tau = 3$. We also note that, when a given axis can be characterized by too much attributes according to their test-values, instead of taking them all, we can restrict the selection by considering only a percentage of the most characteristic ones. i.e, those having the highest absolute test-values. Finally to detect interesting facts in a data cube, for each dimension D_t, we select the following set of characteristic attributes.

$$\Phi_t = \begin{array}{l} a_j^t, \text{ where } \forall\ j \in \{1, \ldots, p_t\}, \\ \exists\ \alpha \in \{1, \ldots, k\} \text{ such as } |V_{\alpha j}^t| \geq 3 \end{array} \tag{2}$$

6. QUALITY OF A DATA REPRESENTATION

In this section, we propose a quality criterion of data cube representation. This criterion measures the homogeneity of geometric distribution of the cube cells. One cell in a data cube contains one or more measures of an OLAP fact. We consider the attributes of a cell in the cube's dimensions as its coordinates in the data space representation. Let $A = (a_{j_1}^1, \ldots, a_{j_t}^t, \ldots, a_{j_d}^d)$ a cell in $\mathcal{C}$, with $t \in \{1, \ldots, d\}$ and $j_t \in \{1, \ldots, p_t\}$. j_t is the index of the attribute that takes the cell A in the dimension D_t. We note $|A|$ the value of the measure contained in the cell A which is equal to NULL in the case where A is empty. For example, in the Figure 3, $|A| = 5.7$ whereas $|Y| =$ NULL. We define the notion of neighborhood of cells as follows.

Definition 1. Let $A = (a_{j_1}^1, \ldots, a_{j_t}^t, \ldots, a_{j_d}^d)$ a cell in a cube $\mathcal{C}$. The cell $B = (b_{j_1}^1, \ldots, b_{j_t}^t, \ldots, b_{j_d}^d)$ is neighbor of A, noted $B \dashv A$, if $\forall t \in \{1, \ldots, d\}$, the coordinates of B satisfy: $b_{j_t}^t = a_{j_t-1}^t$ or $b_{j_t}^t = a_{j_t}^t$ or $b_{j_t}^t = a_{j_t+1}^t$. Except the case where $\forall t \in \{1, \ldots, d\}$ $b_{j_t}^t = a_{j_t}^t$, which means $A = B$.

In Figure 3, the cell B is neighbor of A ($B \dashv A$). Y is also neighbor of A ($Y \dashv A$). Whereas cells S and R are not

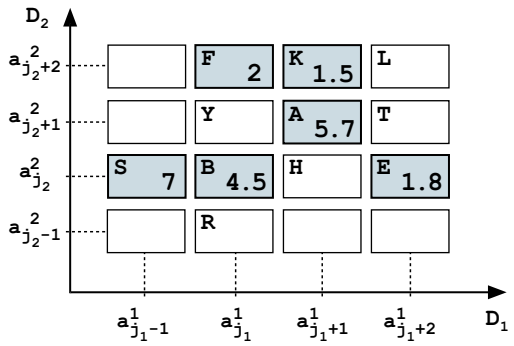


Figure 3: A 2-dimensional example of a data cube.

neighbors of A. This lead us to define the neighborhood of a cell.

Definition 2. Let A a cell of a cube $\mathcal{C}$, we define the neighborhood of A, noted $\mathcal{N}(A)$, by the set of all cells B of $\mathcal{C}$ who are neighbors of A.

$$\mathcal{N}(A) = \{B \in \mathcal{C} \text{ where } B \dashv A\}$$

For example, in Figure 3, the neighborhood of A corresponds to the set $\mathcal{N}(A) = \{F, K, L, T, E, H, B, Y\}$. To quantify similarities between neighbor cells, we define a similarity function δ.

Definition 3. The similarity δ for two cells A and B of the cube $\mathcal{C}$ is defined as follows:

$$\begin{array}{rcl} \delta : \mathcal{C} \times \mathcal{C} & \longrightarrow & \mathbb{R} \\ \delta(A,B) & \longmapsto & \begin{cases} 1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}) & \text{if } A \text{ and } B \text{ are full} \\ 0 & \text{else} \end{cases} \end{array}$$

Where $||A|-|B||$ is the absolute difference of the measures contained in the cells A and B, and $\max(\mathcal{C})$ (respectively, $\min(\mathcal{C})$) is the maximum (respectively, the minimum) measure value in the cube $\mathcal{C}$.

In the cube of the Figure 3, where grayed cells are full and white ones are empty, $\max(\mathcal{C}) = 7$ which matches with the cell S and $\min(\mathcal{C}) = 1.5$ which matches with the cell K. For instance, $\delta(A, B) = 1 - (\frac{|5.7-4.5|}{7-1.5}) \simeq 0.78$ and $\delta(A, Y) = 0$. Let us now introduce the function Δ.

Definition 4. Δ is defined from $\mathcal{C}$ to $\mathbb{R}$ such as:

$$\forall A \in \mathcal{C},\ \Delta(A) = \sum_{B \in \mathcal{N}(A)} \delta(A, B)$$

$\Delta(A)$ corresponds to the sum of the similarities of A with all its full neighbor cells. For instance, in Figure 3, $\Delta(A) = \delta(A, F) + \delta(A, K) + \delta(A, L) + \delta(A, T) + \delta(A, E) + \delta(A, H) + \delta(A, B) + \delta(A, Y) \simeq 1.64$.

Definition 5. We define the crude homogeneity criterion of a data cube $\mathcal{C}$ as:

$$chc(\mathcal{C}) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \sum_{B \in \mathcal{N}(A)} \delta(A, B) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \Delta(A)$$

The crude homogeneity criterion computes the sum of similarities of every couple of full and neighbor cells in a data cube. For instance, in Figure 3, the crude homogeneity criterion is computed as $chc(\mathcal{C}) = \Delta(F) + \Delta(K) + \Delta(A) + \Delta(S) + \Delta(B) + \Delta(E) \simeq 6.67$. Note that, the crude homogeneity criterion of a data cube touches its maximum when all the cells of the cube are fulls and their measures are equals.

$$chc_{max}(\mathcal{C}) = \sum_{A \in \mathcal{C}} \sum_{B \in \mathcal{N}(A)} 1$$

Definition 6. The homogeneity criterion of a data cube is defined as:

$$hc(\mathcal{C}) = \frac{chc(\mathcal{C})}{chc_{max}(\mathcal{C})} = \frac{\sum\limits_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \Delta(A)}{\sum\limits_{A \in \mathcal{C}} \sum\limits_{B \in \mathcal{N}(A)} 1}$$

The homogeneity criterion represent the quality of a multidimensional data representation. This quality is rather better when full and similar cells are neighbors. Indeed, when similar cells are gathered together in specific regions of the space representation of a data cube, this cube is easier to visualize and so a user can directly focus his/her data interpretation on these regions. We summarize the process of computing the homogeneity criterion of an input data cube $\mathcal{C}$ by the algorithm provided in Figure 4. For example, in the Figure 3, $chc_{max}(\mathcal{C}) = 84$, and so the homogeneity criterion of this representation is: $hc(\mathcal{C}) = \frac{6.67}{84} \simeq 0.08$. Nevertheless, such a criterion can not make real sens for a single situation of a data representation. In all cases, we should rather compare it to other data representations of the same cube. In fact, recall that the aim of our method is to organize the facts of an initial data cube representation by arranging attributes in each dimensions according to the order of test-values. Let us note the initial cube $\mathcal{C}_{ini}$ and the organized one $\mathcal{C}_{org}$. To measure the relevance of the organization provided by our method, we compute the gain realized by the homogeneity criterion according to the formula:

$$g = \frac{hc(\mathcal{C}_{org}) - hc(\mathcal{C}_{ini})}{hc(\mathcal{C}_{ini})}$$

We should also note that, for the same cube, its organized representation does not depend of the initial representation because the results of the MCA are insensitive to the order of the input variables.

7. A CASE STUDY

To test and validate our approach, we apply it on a 5-dimensional cube ($d = 5$) that we have constructed from the *Census-Income Database*[2] of the *UCI Knowledge Discovery in Databases Archive*[3]. This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau. The data contains demographic and employment related variables. The constructed cube contains 199 523 facts and one

[2]http://kdd.ics.uci.edu/databases/census-income/census-income.html
[3]http://kdd.ics.uci.edu/

Algorithm MeasureQualityOfCube($\mathcal{C}$)
Input:
$\mathcal{C}$: data cube
Begin
$chc \leftarrow 0$;
$chc_{max} \leftarrow 0$;
for each cell A **in** $\mathcal{C}$ **do**
if ($|A| \neq$ NULL) **then**
for each cell B **in** $\mathcal{N}(A)$ **do**
if ($|B| \neq$ NULL) **then**
$chc \leftarrow chc + (1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}))$;
end if
$chc_{max} \leftarrow chc_{max} + 1$;
end for
else
for each cell B **in** $\mathcal{N}(A)$ **do**
$chc_{max} \leftarrow chc_{max} + 1$;
end for
end if
end for
$hc \leftarrow \frac{chc}{chc_{max}}$;
End

Figure 4: Algorithm MeasureQualityOfCube.

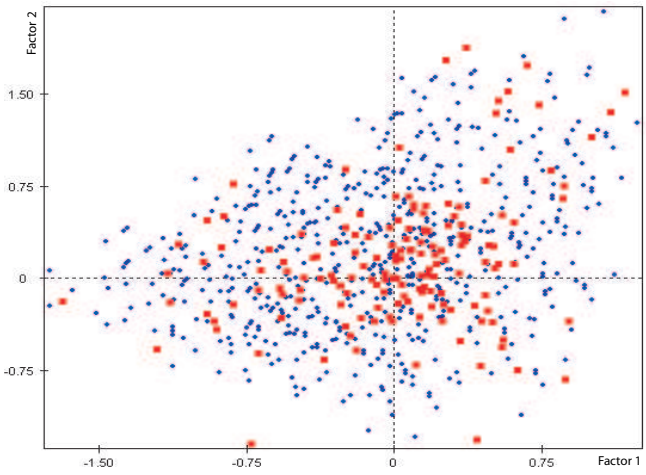


Figure 5: First factorial plane constructed on the Census-Income's data cube.

fact represents a particular profile of a sub population measured by the *Wage per hour*. The Table 2 details the cube's dimensions and measures.

Dimension	p_t
D_1 : Education level	$p_1 = 17$
D_2 : Professional category	$p_2 = 22$
D_3 : State of residence	$p_3 = 51$
D_4 : Household situation	$p_4 = 38$
D_5 : Country of birth	$p_5 = 42$

Table 2: Description of the data cube's dimensions.

According to a binary coding of the cube dimensions, we generate $Z = [Z_1, Z_2, Z_3, Z_4, Z_5]$ as a complete disjunctive table. Z contains 199 523 lines and $p = \sum_{t=1}^{5} p_t = 170$ columns. By applying the MCA on Z we obtain $p - d = 165$ factorial axis F_α. Each axis is associated to an eigenvalue λ_α. Suppose that, according to the histogram of eigenvalues, a user chooses the three first axis ($k = 3$). These axis explain 15.35% of the total inertia of the facts cloud. This contribution does not seem very important at a first sight. But we should note that in a case of a uniform distribution of eigenvalues, we get normally a contribution of $\frac{1}{p-d} = 0.6\%$ per axis, i.e. the three first axis represent an inertia already 25 times more important than a uniform distribution. The Figure 5 displays the first factorial plane we obtain.

The organized *Census-Income* data cube is obtained by sorting the attributes of its dimensions. For each dimension D_t its attributes are sorted by the increasing values of V_{1j}^t, then by V_{2j}^t and then by V_{3j}^t. The Table 3 shows the new attributes' order of the *Professional category* dimension (D_2). Note that j is the index of the original alphabetic order of the attributes. This order is replaced by a new one according to the sort of test-values. In the Figure 6 we can clearly see the visual effect of this arrangement of attributes. This figure displays views of data by crossing the *Professional category* dimension on columns (D_2) and the *Country of birth* dimension on rows (D_5). The representation (a) displays the initial view according to the alphabetic order of attributes, whereas representation (b) displays the same view where attributes are rather sorted according to their test-values. At a first sight, the visual representation (b) is better and more suitable to analyze than (a). This is confirmed by the measure of homogeneity criterion. Indeed, for a sparsity ratio of 63.42%, the homogeneity criterion for the organized cube of representation (b) is $hc(\mathcal{C}_{org}) = 0.134$; whereas it measures $hc(\mathcal{C}_{org}) = 0.112$ for the initial cube of representation (a). i.e, we release a gain $g = 19.64\%$ of homogeneity when arranging the attributes of the cube according to test-values.

		Test-values		
j	Attributes	V_{1j}^1	V_{2j}^1	V_{3j}^1
9	*Hospital services*	-99.90	-99.90	-99.90
14	*Other professional services*	-99.90	-99.90	99.90
17	*Public administration*	-99.90	-99.90	99.90
12	*Medical except hospital*	-99.90	99.90	-99.90
5	*Education*	-99.90	99.90	99.90
7	*Finance insurance*	-99.90	99.90	99.90
19	*Social services*	-99.90	99.90	99.90
8	*Forestry and fisheries*	-35.43	-8.11	83.57
3	*Communications*	-34.05	-99.90	99.90
15	*Personal services except private*	-21.92	-5.50	10.28
13	*Mining*	-6.59	-99.64	-5.25
16	*Private household services*	7.77	51.45	11.68
6	*Entertainment*	40.04	99.90	96.23
1	*Agriculture*	68.66	3.39	-27.38
4	*Construction*	99.90	-99.90	-99.90
10	*Manufact. durable goods*	99.90	-99.90	-99.90
11	*Manufact. nondurable goods*	99.90	-99.90	-99.90
21	*Utilities and sanitary services*	99.90	-99.90	-99.90
22	*Wholesale trade*	99.90	-99.90	-24.37
20	*Transportation*	99.90	-99.90	99.90
18	*Retail trade*	99.90	99.90	-99.90
2	*Business and repair*	99.90	99.90	99.90

Table 3: Attribute's test-values of Professional category dimension.

Furthermore, according to the test of the Equation (2), for each $t \in \{1, \ldots, 5\}$, we select from D_t the set of characteristic attributes for the three selected factorial axis. These characteristic attributes give the best semantic interpration of factorial axis and express strong relationships for their corresponding facts. To avoid great number of possible characteristic attributes per axis, we can consider, for each axis, only the first 50% of attributes having the highest absolute test-values. For instance, in the *Professional category* dimension D_2, the set Φ_2 of characteristic attributes is:

$$\Phi_2 = \left\{ \begin{array}{l} \textit{Hospital services, Other professional services,} \\ \textit{Public administration, Medical except hospital,} \\ \textit{Education, Finance insurance, Social services,} \\ \textit{Forestry and fisheries, Communications,} \\ \textit{Entertainment, Agriculture Construction,} \\ \textit{Manufact. durable goods,} \\ \textit{Manufact. nondurable goods,} \\ \textit{Utilities and sanitary services, Wholesale trade,} \\ \textit{Transportation, Retail trade,} \\ \textit{Business and repair services} \end{array} \right\}$$

In the same way, we apply the test of the Equation (2) on the other dimensions of the cube. In the representation (b) of the Figure 6, we clearly see that the zones of facts corresponding to characteristic attributes of the dimensions D_2 and D_5 seem to be interesting and denser than other regions in the data space representation. These zones contains relevant information and reflect interesting association between facts. For example, we can easily note that industrial and physical jobs, like construction, agriculture and manufacturing are highly performed by native Latin Americans from Ecuador, Peru, Nicaragua and Mexico for example. At the opposite, Asian people from India, Iran, Japan and China are rather concentrated in commerce and trade.

8. EXPERIMENTAL RESULTS

We have realized some experiments of our approach in order to appreciate its efficiency. To achieve this, we based our experiments on the *Census-Income* data cube presented in section 7. The aim of these experiments is to measure the homogeneity gain realized by our MCA-based organization method on data representations with different sparsity ratios. To vary sparsity we proceed by a random sampling of the data set of the initial n facts of the considered cube.

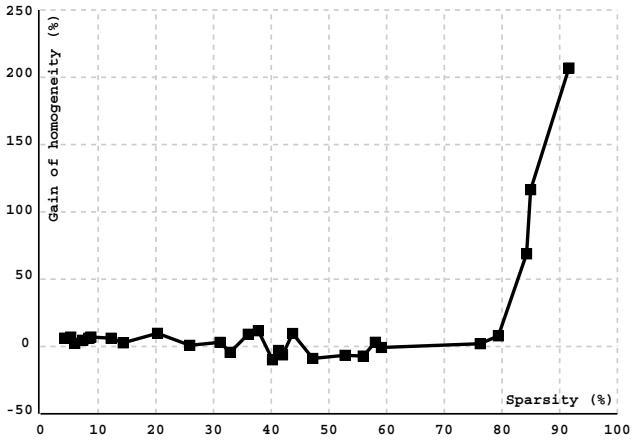


Figure 7: Evolution of the homogeneity gain according to sparsity.

According to the Figure 7, the homogeneity gain has an increasing general trend. Nevertheless, we should note that for low sparsity ratios, the curve is rather oscillating around the null value of the homogeneity gain. In fact, when sparsity is less then 60%, the gain does not have a constant variation. It sometimes drops to negative values. This means that our method does not bring a value added to the quality of data representation. For dense data cubes, the employment of our method is not always significant. This is naturally due to the construction of the homogeneity criterion which closely depends of the number of empty and full cells. It can also be due to the structure of the random data samples that can generate data representations already having good qualities and high homogeneity values.

Our MCA-based organization method is rather interesting for data representations with high sparsity. In the Figure 7, we clearly see that curve is rapidly increasing to high positive values of gain when sparsity is greater thant 60%. Actually, with high relative number of empty cells in a data cube, we have a large manoeuvre margin for concentrating similar full cells and gathering them in the space representation. This shows the vocation of using our approach in order to enhance the visual quality representation, and thus the analysis of huge and sparse data cubes.

9. CONCLUSION AND FUTURE WORK

In this paper we have introduced a MCA-based approach to enhance the space representation of large and sparse data cubes. This approach aims to provide an assistance to the OLAP user and helps him/her to easily explore huge volumes of data. For a given data cube, we compute the test-values of its attributes. According to these test-values, we arrange attributes of each dimension and so display in an appropriate way the space representation of facts. This representation provides better property for data visualization since it gather full cells expression interesting relationships of data. We also identify relevant regions of facts in this data representation by detecting characteristic attributes of factorial axis. This solve the problem of high dimensionality and sparsity of data and allows the user to directly focus his exploration and data interpretation on these regions.

We have also proposed an homogeneity criterion to measure the quality of data representations. This criterion is based on the notion of geometric neighborhood of cells and their measures' similarities. Through experiments we led on real world data, our criterion proved the efficiency of our approach for huge and sparse data cubes.

Currently, we are studying some possible extensions for this work. For instance, we are addressing the problem of materializing organized representations. We are also trying to envolve our approach in order to make it able to take into account the data updates.

10. REFERENCES

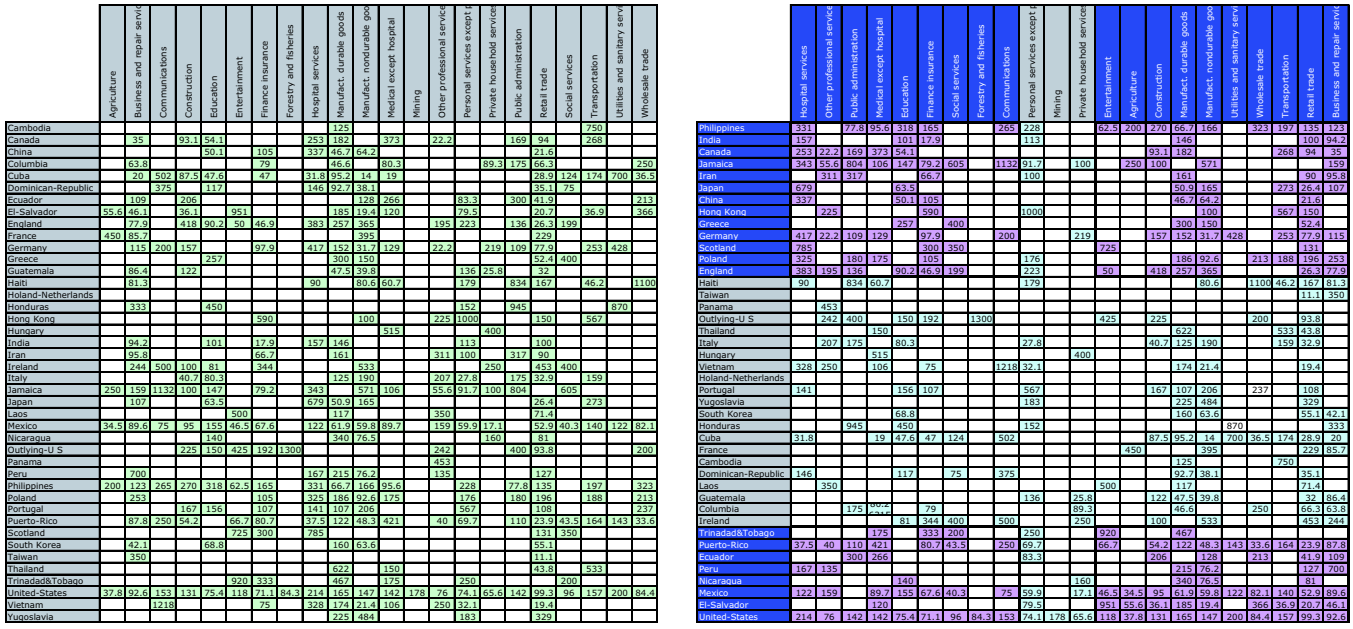
[1] D. Barbará and M. Sullivan. Quasi-Cubes: Exploiting Approximations in Multidimensional Databases. *SIGMOD Record*, 26(3):12–17, 1997.

[2] R. Cattell. The scree test for the number of fators. *Multivariate Behavioral Research*, 1:245–276, 1966.

[3] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Record*, 26(1):65–74, 1997.

[4] Y. W. Choong, A. Laurent, D. Laurent, and P. Maussion. Résumé de cube de données multidimensionnelles à l'aide de règles floues. In R. des Nouvelles Technologies de l'Information, editor, *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, volume 1, pages 95–106, Clermont-Ferrand, France, January 2004.

[5] Y. W. Choong, D. Laurent, and P. Marcel. Computing Appropriate Representations for Multidimensional



	Agriculture	Business and repair servic	Communications	Construction	Education	Entertainment	Finance insurance	Forestry and fisheries	Hospital services	Manufact. durable goods	Manufact. nondurable goo	Medical except hospital	Mining	Other professional service	Personal services except	Private household services	Public administration	Retail trade	Social services	Transportation	Utilities and sanitary serv	Wholesale trade
Cambodia										125										750		
Canada		35		93.1	54.1				253	182		373		22.2			169	94		268		
China					50.1		105		337	46.7	64.2							21.6				
Columbia		63.8					79			46.6		80.3				89.3	175	66.3				250
Cuba		20	502	87.5	47.6		47		31.8	95.2	14	19						28.9	124	174	700	36.5
Dominican-Republic			375		117				146	92.7	38.1							35.1	75			
Ecuador		109		206							128	266			83.3		300	41.9				213
El-Salvador	55.6	46.1		36.1		951				185	19.4	120			79.5			20.7		36.9		366
England		77.9		418	90.2	50	46.9		383	257	365			195	223		136	26.3	199			
France	450	85.7									395							229				
Germany		115	200	157			97.9		417	152	31.7	129		22.2		219	109	77.9		253	428	
Greece					257					300	150							52.4	400			
Guatemala		86.4		122						47.5	39.8				136	25.8		32				
Haiti		81.3							90		80.6	60.7			179		834	167		46.2		1100
Holand-Netherlands																						
Honduras		333			450						571	106			152		945		605		870	
Hong Kong							590			100				225	1000			150		567		
Hungary												515				400						
India		94.2			101		17.9		157	146					113			100				
Iran		95.8					66.7			161				311	100		317	90				
Ireland		244	500	100	81		344				533					250		453	400			
Italy				40.7	80.3					125	190			207	27.8		175	32.9		159		
Jamaica	250	159	1132	100	147		79.2		343		571	106		55.8	91.7	100	804		605			
Japan		107			63.5				679	50.9	165							26.4		273		
Laos						500				117				350				71.4				
Mexico	34.5	89.6	75	95	155	46.5	67.6		122	61.9	59.8	89.7		159	59.9	17.1		52.9	40.3	140	122	82.1
Nicaragua					140					340	76.5					160		81				
Outlying-U S				225	150	425	192	1300						242			400	93.8				200
Panama														453								
Peru		700							167	215	76.2			135				127				
Philippines	200	123	265	270	318	62.5	165		331	66.7	166	95.6			228		77.8	135		197		323
Poland		253					105		325	186	92.6	175			176		180	196		188		213
Portugal				167	156		107		141	107	206				567			108				237
Puerto-Rico		87.8	250	54.2		66.7	80.7		37.5	122	48.3	421		40	69.7		110	23.9	43.5	164	143	33.6
Scotland						725	300		785									131	350			
South Korea		42.1			68.8					160	63.6							55.1				
Taiwan		350																11.1				
Thailand										622		150						43.8		533		
Trinadad&Tobago						920	333			467		175			250				200			
United-States	37.8	92.6	153	131	75.4	118	71.1	84.3	214	165	147	142	178	76	74.1	65.6	142	99.3	96	157	200	84.4
Vietnam			1218				75		328	174	21.4	106		250	32.1			19.4				
Yugoslavia										225	484				183			329				

(a)

	Hospital services	Other professional service	Public administration	Medical except hospital	Education	Finance insurance	Social services	Forestry and fisheries	Communications	Personal services except	Mining	Private household services	Entertainment	Agriculture	Construction	Manufact. durable goods	Manufact. nondurable goo	Utilities and sanitary serv	Wholesale trade	Transportation	Retail trade	Business and repair servic
Philippines	331		77.8	95.6	318	165			265	228			62.5	200	270	66.7	166		323	197	135	123
India	157				101	17.9				113						146					100	94.2
Canada	253	22.2	169	373	54.1										93.1	182				268	94	35
Jamaica	343	55.6	804	106	147	79.2	605		1132	91.7		100		250	100		571					159
Iran		311	317		66.7					100						161					90	95.8
Japan	679			63.5												50.9	165			273	26.4	107
China	337			50.1	105											46.7	64.2				21.6	
Hong Kong		225			590				1000								100			567	150	
Greece				257			400									300	150				52.4	
Germany	417	22.2	109	129		97.9			200			219			157	152	31.7	428		253	77.9	115
Scotland	785				300	350							725								131	
Poland	325		180	175		105				176						186	92.6		213	188	196	253
England	383	195	136		90.2	46.9	199			223			50		418	257	365				26.3	77.9
Haiti	90		834	60.7						179							80.6		1100	46.2	167	81.3
Taiwan																					11.1	350
Panama		453																				
Outlying-U S		242	400		150	192		1300					425		225				200		93.8	
Thailand				150												622				533	43.8	
Italy		207	175		80.3					27.8					40.7	125	190			159	32.9	
Hungary				515								400										
Vietnam	328	250		106		75			1218	32.1						174	21.4				19.4	
Holand-Netherlands																						
Portugal	141				156	107				567					167	107	206		237		108	
Yugoslavia										183						225	484				329	
South Korea					68.8											160	63.6				55.1	42.1
Honduras			945		450					152								870				333
Cuba	31.8			19	47.6	47	124		502						87.5	95.2	14	700	36.5	174	28.9	20
France														450			395				229	85.7
Cambodia																125				750		
Dominican-Republic	146				117		75		375							92.7	38.1				35.1	
Laos		350											500			117					71.4	
Guatemala										136		25.8			122	47.5	39.8				32	86.4
Columbia			175		79							89.3				46.6			250		66.3	63.8
Ireland				81	344	400			500			250			100		533				453	244
Trinadad&Tobago			175		333	200				250			920			467						
Puerto-Rico	37.5	40	110	421		80.7	43.5		250	69.7			66.7		54.2	122	48.3	143	33.6	164	23.9	87.8
Ecuador			300	266						83.3					206		128		213		41.9	109
Peru	167	135														215	76.2				127	700
Nicaragua					140							160				340	76.5				81	
Mexico	122	159		89.7	155	67.6	40.3		75	59.9		17.1	46.5	34.5	95	61.9	59.8	122	82.1	140	52.9	89.6
El-Salvador				120						79.5			951	55.6	36.1	185	19.4		366	36.9	20.7	46.1
United-States	214	76	142	142	75.4	71.1	96	84.3	153	74.1	178	65.6	118	37.8	131	165	147	200	84.4	157	99.3	92.6

(b)

Figure 6: (a) Initial and (b) organized data representations of the Census-Income cube.

Data. *Data & knowledge Engineering Journal*, 45(2):181–203, 2003.

[6] B. Escofier and B. Leroux. Etude de trois problèmes de stabilité en analyse factorielle. *Publication de l'institut Statistique de l'Université de Paris*, 11:1–48, 1972.

[7] J. Feng, Q. Fang, and H. Ding. PrefixCube: Prefix-sharing Condensed Data Cube. In *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP (DOLAP 04)*, pages 38–47, Washington D.C., U.S.A., November 2004.

[8] Y. Feng, D. Agrawal, A. E. Abbadi, and A. Metwally. Range CUBE: Efficient Cube Computation by Exploiting Data Correlation. In *Proceedings of the 20th International Conference on Data Engineering*, pages 658–670, 2004.

[9] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.

[10] J.P. Benzecri. *Correspondence Analysis Handbook*. Marcel Dekker, hardcover edition, January 1992.

[11] H. Kaiser. A note on Guttman's lower bound for the number of common factors. *Brit. J. Statist. Psychol.*, 14:1–2, 1961.

[12] R. Kimball. *The Data Warehouse toolkit*. John Wiley & Sons, 1996.

[13] L. V. Lakashmanan, J. Pei, and J. Han. Quotient Cube: How to Summarize the Semantics of a Data Cube. In *Proceedings of International Conference of Very Large Data Bases, VLDB'02*, 2002.

[14] L. V. Lakshmanan, J. Pei, and Y. Zhao. QC-Trees: An Efficient Summary Structure for Semantic OLAP. In A. Press, editor, *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 64–75, 2003.

[15] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunold, Paris, 3^e édition edition, 2000.

[16] X. Li, J. Han, and H. Gonzalez. High-Dimensional OLAP: A Minimal Cubing Approach. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004)*, pages 528–539, August 2004.

[17] E. Malinvaud. Data analysis in applied socio-economic statistics with special consideration of correspondence analysis. In *Proceedings of Marketing Science Conference*, Jouy en Josas, France, 1987.

[18] K. A. Ross and D. Srivastava. Fast Computation of Sparse Datacubes. In *Proceedings of the 23rd International Conference of Very Large Data Bases, VLDB'97*, pages 116–125. Morgan Kaufmann, 1997.

[19] J. Shanmugasundaram, U. M. Fayyad, and P. S. Bradley. Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 223–232, August 1999.

[20] Y. Sismanis, A. Deligiannakis, N. Roussopoulos, and Y. Kotidis. Dwarf: Shrinking the PetaCube. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 464–475. ACM Press, 2002.

[21] J. S. Vitter and M. Wang. Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of Data*, pages 193–204, Philadelphia, Pennsylvania, U.S.A., June 1999. ACM Press.

[22] W. Wang, H. Lu, J. Feng, and J. X. Yu. Condensed Cube: An Effective Approach to Reducing Data Cube Size. In *Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE'02)*, 2002.