Efficient Multidimensional Data Representations Based on Multiple Correspondence Analysis

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

In the On Line Analytical Processing (OLAP) context, exploration of huge and sparse data cubes is a tedious task which does not always lead to efficient results. In this paper, we couple OLAP with the Multiple Correspondence Analysis (MCA) in order to enhance visual representations of data cubes and thus, facilitate their interpretations and analysis. We also provide a quality criterion to measure the relevance of obtained representations. The criterion is based on a geometric neighborhood concept and a similarity metric between cells of a data cube. Experimental results on real data proved the interest and the efficiency of our approach.

Categories and Subject Descriptors

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

1. INTRODUCTION

On Line Analytical Processing (OLAP) is a technology supported by most data warehousing systems [3]. It provides a platform for analyzing data according to multiple dimensions and multiple hierarchical levels. Data are presented in multidimensional views, commonly named data cubes. A data cube can be considered as a space representation composed by a set of cells. Each cell represents a precise fact associated with one or more measures and identified by coordinates represented by one attribute from each dimension. OLAP provides users with visual based tools to summarize, explore and navigate into data cubes in order to detect interesting and relevant information. However, exploring data cubes is not always an easy task to perform. Obviously, in large cubes with sparse data, the whole analysis process becomes tedious and complex.

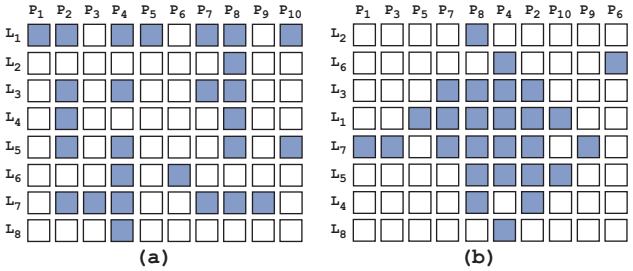

Figure 1: Example of different representations of a 2-dimensional data cube.

For instance, consider the cube of Figure 1 which displays sales of products $(P_1, \ldots, P_{10})$ crossed by geographic locations of stores $(L_1, \ldots, L_8)$. On the one hand, in representation 1(a), full cells (gray cells) are displayed randomly according to a lexical order of *attributes* – also named *members* – in each dimension. The way the cube is displayed does not provide an attractive representation that visually helps to interpret data. On the other hand, Figure 1(b) contains the same information as Figure 1(a). However, it displays a data representation which is visually easier to analyze. Figure 1(b) gathers full cells together and separates them from empty ones. Such a representation is naturally more comfortable and enables easy and efficient analysis. Note that representation (b) can be interactively constructed from representation (a) via some traditional OLAP operators. However, this suppose that the user intuitively knows how to arrange attributes of each dimension. We propose an automatic identification and an arrangement of interesting facts. Our a method enables to get relevant facts expressing relationships and displays them in an appropriate way in order to enhance the exploration process independently of the cube's size. In order to do so, we carry out a Multiple Correspondence Analysis [2] (MCA) on a data cube. Basically, MCA is a powerful describing method even for huge volumes

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

of data. It factors categorical variables and displays data in a factorial space constructed by orthogonal system of axes which provides relevant views of data. We focus on relevant OLAP facts associated with characteristic attributes (variables) provided by factorial axes. These facts are interesting since they reflect relationships and concentrate significant information. In order to ensure an appropriate representation of these facts, we highlight them and arrange their attributes in the data space representation by using *test-values* [4]. We also propose a novel criterion to measure the homogeneity of cells' distribution in the space representation of a data cube. This criterion is based on a concept of geometric neighborhood of cells. It also takes into account a similarity metric of cells' measures and therefore provides a scalar quantification for the homogeneity of a given data cube representation.

2. OVERVIEW OF OUR METHOD

Our method can be directly applied on a data cube $\mathcal{C}$ or on a data view (a sub-cube) extracted from $\mathcal{C}$. 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 to analyse. In order to lighten notations, we assume that a user has selected a data cube $\mathcal{C}$, with d dimensions $(D_t)_{1\leq t\leq d}$, m measures $(M_q)_{1\leq q\leq m}$ 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 dimension 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 method 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, we establish a new arrangement of its attributes a_j^t. This arrangement displays multidimensional information in a more appropriate manner. In a second step, our method detects from the resulted representation relevant facts expressing interesting relationships. In order to do so, we select from each dimension D_t a subset Φ_t of significant attributes, also named characteristic attributes. The crossing of these particular attributes allows to identify relevant cells in the cube.

We base our method on the MCA [2], which is a factorial technique that displays categorical variables in a property space and maps their associations in two or more dimensions. From a table of n observations and p categorical variables ($p < n$), the MCA provides orthogonal axes 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. 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* 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.

3. APPLYING THE MCA ON A DATA CUBE

Like all statistical techniques, the MCA needs a tabular representation of input data. Therefore, we can not apply it directly on a multidimensional representation. We need to convert $\mathcal{C}$ to a *complete disjunctive table*. The conversion consists in transforming each dimension D_t into a binary matrix Z_t with n rows and p_t columns. The i^{th} row 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:

$$z_{ij}^t = \begin{cases} 1 & \text{if the fact } i \text{ takes the attribute } a_j^t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

By merging the d matrices Z_t, we obtain a complete disjunctive table $Z = [Z_1, Z_2, \ldots, Z_t, \ldots, Z_d]$ with n rows and p columns. It 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 obtain a very huge matrix Z. Once the complete disjunctive table Z is built, the MCA starts by constructing a matrix $B = Z'Z$ – called *Burt* table –, where Z' is the transposed matrix of Z. *Burt* table B is a (p, p) symmetric matrix which contains all the category marginal on the main diagonal and all possible cross-tables of the d dimensions of $\mathcal{C}$ in the off-diagonal. Let X be a (p, p) diagonal matrix which has the same diagonal elements of B and zeros otherwise. We construct from Z and X a new matrix $S = \frac{1}{d}\ Z'ZX^{-1} = \frac{1}{d}\ BX^{-1}$

By diagonalizing S, we obtain $(p - d)$ diagonal elements, called *eigenvalues* and denoted λ_α. Each eigenvalue λ_α is associated to a directory vector u_α and corresponds to a factorial axis F_α, where $Su_\alpha = \lambda_\alpha u_\alpha$. 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, we only keep the first, two or three axes of inertia [5, 1]. In [2], Benzecri suggests that the number k of axes to keep should be fixed by user's capacity to give them a meaningful interpretation. It is 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.

4. ORGANIZING DATA CUBES AND DETECTING RELEVANT FACTS

Usually in a factorial analysis, relative contributions of variables are used to give sense to the axes. A relative contribution shows the percentage 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 to 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 [4]. In the followings, we present the theoretical principle of test-values applied to the context of our approach.

4.1 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). $\varphi_{\alpha j}^t = \frac{1}{n_j^t \sqrt{\lambda_\alpha}} \sum_{i \in I(a_j^t)} \psi_{\alpha i}$ is 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 $Y^t_{\alpha j} = \frac{1}{n^t_j} \sum_{i \in I(a^t_j)} \psi_{\alpha i}$, where $\mathrm{E}(Y^t_{\alpha j}) = 0$ and $\mathrm{VAR}_{H_0}(Y^t_{\alpha j}) = \frac{n-n^t_j}{n-1} \frac{\lambda_\alpha}{n^t_j}$. Note that $\varphi^t_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} Y^t_{\alpha j}$. Thus, $\mathrm{E}(\varphi^t_{\alpha j}) = 0$, and $\mathrm{VAR}_{H_0}(\varphi^t_{\alpha j}) = \frac{n-n^t_j}{n-1} \frac{1}{n^t_j}$. Therefore, the test-value of the attribute a^t_j is:

$$V^t_{\alpha j} = \sqrt{n^t_j \frac{n-1}{n-n^t_j}} \varphi^t_{\alpha j} \quad (2)$$

$V^t_{\alpha j}$ measures the number of standard deviations between the attribute a^t_j (the gravity center of the n^t_j facts) and the center of the 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.

4.2 Arrangement of attributes

In traditional representation of data cubes, attributes are usually organized according to a lexical order such as alphabetic order for a *geographic* dimension or chronological order for a *time* dimension. In a formal way, we consider that attributes of a dimension D_t are geometrically organized in a cube representation according to the order of indices j_t. i.e, the attribute $a^t_{j_t-1}$ precedes $a^t_{j_t}$, and $a^t_{j_t}$ precedes $a^t_{j_t+1}$, and so on (see the example of Figure 2). We propose to exploit the test-values of attributes in order to organize differently the cube's facts. Especially for large and sparse cubes, this new organization displays a relevant data representation suitable for analysis. In order to do so, for each dimension, we sort its attributes a^t_j according to the increasing order of their k first test-values $V^t_{\alpha j}$ on axes F_α. Thus, we obtain a new order of indices j, which provides a new arrangement of attributes a^t_j in each dimension D_t.

4.3 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 roughly corresponds to an error threshold of 5%. 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 performing p tests for all attributes. To minimize this accumulation of errors, we 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 only considering a percentage of the most characteristic ones. Thus, for each dimension D_t, we select the following set of characteristic attributes:

$$\Phi_t = \left\{ \begin{array}{l} a^t_j, \text{ where } \forall\ j \in \{1, \ldots, p_t\}, \\ \exists\ \alpha \in \{1, \ldots, k\} \text{ such as } |V^t_{\alpha j}| \geq 3 \end{array} \right\} \quad (3)$$

5. QUALITY OF REPRESENTATIONS

We propose a quality criterion of data cube representations which measures the homogeneity of geometric distribution of cells. Attributes of a cell represent its coordinates according to dimensions of the data space representation. Let $A = (a^1_{j_1}, \ldots, a^t_{j_t}, \ldots, a^d_{j_d})$ be a cell in $\mathcal{C}$. j_t is the index of the attribute taken by A in dimension D_t. We assume that $|A|$ is the value of the measure contained in A, which is equal to NULL if A is empty. For example, in Figure 2, $|A| = 5.7$ and $|Y| = \mathsf{NULL}$.

Let $B = (b^1_{j_1}, \ldots, b^t_{j_t}, \ldots, b^d_{j_d})$ be a second cell in $\mathcal{C}$. B is said neighbor of A, noted $B \dashv A$, if $\forall t \in \{1, \ldots, d\}$, the coordinates of B satisfy: $b^t_{j_t} = a^t_{j_t-1}$ or $b^t_{j_t} = a^t_{j_t}$ or $b^t_{j_t} = a^t_{j_t+1}$. This definition does not include the case where $\forall t \in \{1, \ldots, d\}\ b^t_{j_t} = a^t_{j_t}$, which corresponds to the situation where $A = B$. For example, in Figure 2, 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 neighbors of A.

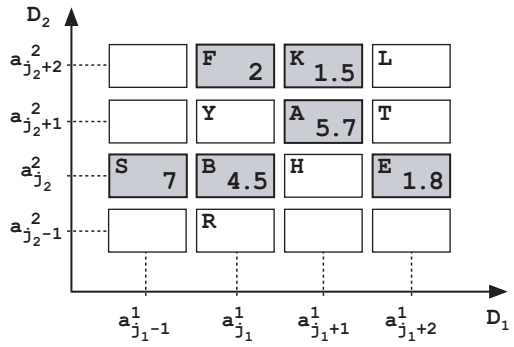

Figure 2: A 2-dimensional example of a data cube.

The neighborhood of A, noted $\mathcal{N}(A)$, defines the set of all cells B of $\mathcal{C}$ neighbors of A. For example, in Figure 2, the neighborhood of A corresponds to the set $\mathcal{N}(A)$={*F, K, L, T, E, H, B, Y*}. In a formal notation:

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

We also define a similarity metric δ of two cells A and B from a cube $\mathcal{C}$ according to the following function:

$$\begin{array}{lcl} \delta : \mathcal{C} \times \mathcal{C} & \longrightarrow & \mathbb{R} \\ \delta(A,B) & \longmapsto & \left\{ \begin{array}{ll} 1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}) & \text{if } A \text{ and } B \text{ are full} \\ 0 & \text{otherwise} \end{array} \right. \end{array}$$

where $||A| - |B||$ is the absolute difference of the measures contained in A and B, and $\max(\mathcal{C})$ (respectively, $\min(\mathcal{C})$) is the maximum (respectively, the minimum) existant measure value in $\mathcal{C}$. In Figure 2, where grayed cells are full and white ones are empty, $\max(\mathcal{C}) = 7$, which corresponds to the cell S, and $\min(\mathcal{C}) = 1.5$, which corresponds to the cell K. For instance, $\delta(A,B) = 1 - (\frac{|5.7-4.5|}{7-1.5}) \simeq 0.78$, and $\delta(A,Y) = 0$.

We introduce now the metric Δ defined from $\mathcal{C}$ to $\mathbb{R}$ such as $\forall A \in \mathcal{C},\ \Delta(A) = \sum_{B \in \mathcal{N}(A)} \delta(A,B)$. It corresponds to the sum of the similarities of A with all its full neighbor cells. For example, in Figure 2, $\Delta(A) = \delta(A,F) + \delta(A,K) + \delta(A,L) + \delta(A,T) + \delta(A,E) + \delta(A,H) + \delta(A,B) + \delta(A,Y) \simeq 1.64$.

Therefore, we can define the crude homogeneity criterion of a data cube $\mathcal{C}$ as:

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

This criterion computes the sum of similarities of every couple of full and neighbor cells in a data cube $\mathcal{C}$. In Figure 2, 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 value $chc_{max}(\mathcal{C})$ when all cells of $\mathcal{C}$ are full and have the same measure value. Therefore, we consider that $chc_{max}(\mathcal{C}) = \sum_{A \in \mathcal{C}} \sum_{B \in \mathcal{N}(A)} 1$. Finally, we define the homogeneity criterion of a data cube as follows:

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

The homogeneity criterion represents 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. One user can therefore directly focus his analysis on these regions. Nevertheless, such a criterion can not make real sense for a single data representation. We should rather compare it to other representations of the same cube. Recall also that we aim at organizing facts of an initial data cube representation by arranging attributes in each dimensions. Let $\mathcal{C}_{ini}$ be the initial cube representation, and $\mathcal{C}_{org}$ be the organized one. To measure the relevance of the organization provided by our method, we compute its realized gain of homogeneity:

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

6. CASE STUDY

We apply our method on a 5-dimensional cube ($d = 5$) that we constructed from the *Census-Income Database*[1] of the *UCI Knowledge Discovery in Databases Archive*[2]. 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. One fact represents a particular profile of a sub population measured by the *Wage per hour*. The following table illustrates the cube's dimensions.

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	$p_5 = 42$

According to a binary coding of the cube dimensions, we generate a complete disjunctive table $Z = [Z_1, Z_2, Z_3, Z_4, Z_5]$. Z contains 199 523 rows and $p = \sum_{t=1}^{5} p_t = 170$ columns. By applying the MCA on Z we obtain $p - d = 165$ factorial axes 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 axes ($k = 3$). These axes 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 normally get a contribution of $\frac{1}{p-d} = 0.6\%$ per axis, i.e. the three first axes represent an inertia already 25 times more important than a uniform distribution.

For each dimension D_t of the *Census-Income* data cube, its attributes are sorted according to the increasing values of V^t_{1j}, then by V^t_{2j}, and then by V^t_{3j}. Table 1 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 Figures 3(a) and 3(b), we can clearly see the visual effect of this new arrangement of attributes. These figures display views of data by crossing the *Professional category* dimension on columns (D_2) and the *Country of birth* dimension on rows (D_5). Representation 3(a) displays the initial view according to the alphabetic order of attributes, whereas representation 3(b) displays the same view where attributes are rather sorted according to their test-values.

		Test-values		
j	Attributes	V^1_{1j}	V^1_{2j}	V^1_{3j}
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 1: Attribute's test-values of *Professional category* dimension.

We emphasize that our method does not cope with compressing dimensions of a data cube. We do not also aim at decreasing the sparsity of a data cube. Nevertheless, we act on this sparsity and reduce its negative effect on OLAP interpretation. We rather arrange differently original facts within a visual effect that gathers them as well as possible in the space representation of the data cube. At a first sight, representation 3(b) is more suitable to interpretation than 3(a). We clearly distinguish in Figure 3(b) four dense regions of full cells. In these regions, neighbor cells are more homogeneous than in the rest of the space representation. This result is confirmed by the homogeneity criterion. Indeed, for a sparsity ratio of 63.42%, the homogeneity criterion of the organized cube in representation 3(b) is $hc(\mathcal{C}_{org}) = 0.17$; whereas it measures $hc(\mathcal{C}_{ini}) = 0.14$ for the initial cube in representation 3(a), i.e, our method enables a gain of homogeneity $g = 17.19\%$.

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

[2] http://kdd.ics.uci.edu/

	Agriculture	Business and repair services	Communications	Construction	Education	Entertainment	Finance insurance	Forestry and fisheries	Hospital services	Manufact. durable goods	Manufact. nondurable goods	Medical except hospital	Mining	Other professional services	Personal services except private	Private household services	Public administration	Retail trade	Social services	Transportation	Utilities and sanitary services	Wholesale trade
Cambodia										125.0										750.0		
Canada		35.0		93.1	54.1			112.5	253.1	182.3		373.4		22.2			169.2	94.0		267.6	11.1	350.0
China	622.0			40.7	50.1		105.0	566.7	336.8	46.7	64.2	60.7					833.8	21.6			329.0	206.3
Columbia							79.0			46.6		80.3					175.0					
Cuba			501.5						31.8			19.0						28.9				
Dominican-Republic			375.0		116.7				146.0	92.7	38.1							35.1	75.0			
Ecuador	107.2	109.1	250.0	205.6	515.0		206.7	68.8			128.1	265.6	100.0				300.0	41.9	175.0		333.3	212.5
El-Salvador	55.6	46.1		36.1	81.0	950.8	344.0			184.7	19.4	120.0			79.5			20.7	400.0	36.9		365.6
England		77.9	222.7	418.1	90.2	50.0	46.9		383.0	257.1	365.0			194.7			136.4	26.3	198.9			
France	450.0										394.8							229.0				
Germany		115.0	200.0	157.1			97.9		417.2	152.3	31.7	128.6		22.2		218.9	108.7	77.9		253.1	428.2	
Greece					257.1					300.0	150.0			241.7			400.0	52.4	400.0			63.6
Guatemala				121.8						47.5	39.8				136.2	25.8						
Haiti									90.0		80.6				178.7							###
Holand-Netherlands											21.4											
Honduras															151.7		945.0					
Hong Kong	125.4			190.5			590.4	183.3			100.0			225.0	###			150.0		566.7	55.1	484.3
Hungary																400.0						
India		94.2			101.2		17.9	228.1	157.2	145.9								100.0			167.1	81.3
Iran		95.8		225.0			66.7			160.7				311.1	100.0		316.7	90.0		159.0		
Ireland			500.0	100.0							533.3											
Italy					80.3										27.8			32.9				
Jamaica	250.0	158.8	###	100.0	147.0		79.2		343.1		571.4	106.0		55.6	91.7	100.0	803.8		604.7	533.3	19.4	
Japan		107.1			63.5	425.0		192.1	678.9	50.9	164.6							26.4	150.0	273.3	107.5	
Laos						500.0				116.6				350.0				71.4				
Mexico	34.5	89.6	75.0	95.0	155.2	46.5	67.6		122.2	61.9	59.8	89.7		159.1	59.9	17.1		52.9	40.3	140.3	121.7	82.1
Nicaragua	159.5		83.3		140.0			47.6		340.0	76.5	65.6		74.1		160.0	178.3	81.0			85.7	
Outlying-U S								###										93.8				200.0
Panama														452.5								
Peru	225.0	699.6	69.7		106.3		47.0	450.0	166.7	215.4	76.2			134.5				127.3	124.2	86.4	20.0	32.0
Philippines	200.0	122.7	265.0	270.0	317.8	62.5	165.0		331.1	66.7	166.1	95.6					77.8	134.7		197.3		322.7
Poland		252.9	175.6				105.0		325.0	185.5	92.6	175.2					180.0	196.2		187.5		212.5
Portugal				166.7	155.6		107.1		141.1													236.7
Puerto-Rico		87.8	250.0	54.2		66.7	80.7	250.0	37.5	122.3	48.3	420.7		40.0			110.1	23.9	43.5	163.8	142.9	33.6
Scotland				87.5		725.0	300.0		785.0	95.2	14.0		23.9					131.3	350.0	173.6	700.0	36.5
South Korea																					870.0	
Taiwan																				46.2		
Thailand												150.0						43.8				
Trinadad&Tobago	66.3	243.8		63.8		920.0	333.3	89.3		466.7		175.0						453.0	200.0			250.0
United-States	37.8	92.6	153.4	130.6	75.4	117.9	71.1	84.3	214.4	165.4	146.9	141.7		76.0			142.1	99.3	96.0	157.0	199.9	84.4
Vietnam			###				75.0		327.5	173.8				250.0	32.1							
Yugoslavia		42.1																				

(a)

	Hospital services	Other professional services	Public administration	Medical except hospital	Education	Finance insurance	Social services	Forestry and fisheries	Communications	Personal services except private	Mining	Private household services	Entertainment	Agriculture	Construction	Manufact. durable goods	Manufact. nondurable goods	Utilities and sanitary services	Wholesale trade	Transportation	Retail trade	Business and repair services
Philippines	331.1		77.8	95.6	317.8	165.0			265.0				62.5	200.0	270.0	66.7	166.1		322.7	197.3	134.7	122.7
India	157.2				101.2	17.9		228.1								145.9		167.1	81.3		100.0	94.2
Canada	253.1	22.2	169.2	373.4	54.1			112.5							93.1	182.3		11.1	350.0	267.6	94.0	35.0
Jamaica	343.1	55.6	803.8	106.0	147.0	79.2	604.7		###	91.7		100.0		250.0	100.0		571.4	19.4		533.3		158.8
Iran		311.1	316.7			66.7				100.0					225.0	160.7				159.0	90.0	95.8
Japan	678.9				63.5		150.0	192.1					425.0			50.9	164.6	107.5		273.3	26.4	107.1
China	336.8		833.8	60.7	50.1	105.0		566.7						622.0	40.7	46.7	64.2	329.0	206.3		21.6	
Hong Kong		225.0				590.4		183.3		###				125.4	190.5		100.0	55.1	484.3	566.7	150.0	
Greece		241.7	400.0		257.1		400.0									300.0	150.0		63.6		52.4	
Germany	417.2	22.2	108.7	128.6		97.9			200.0			218.9			157.1	152.3	31.7	428.2		253.1	77.9	115.0
Scotland	785.0					300.0	350.0				23.9		725.0		87.5	95.2	14.0	700.0	36.5	173.6	131.3	
Poland	325.0		180.0	175.2		105.0			175.6							185.5	92.6		212.5	187.5	196.2	252.9
England	383.0	194.7	136.4		90.2	46.9	198.9		222.7				50.0		418.1	257.1	365.0				26.3	77.9
Haiti	90.0									178.7							80.6		###			
Taiwan																				46.2		
Panama		452.5																				
Outlying-U S								###											200.0		93.8	
Thailand				150.0																	43.8	
Italy					80.3					27.8											32.9	
Hungary												400.0										
Vietnam	327.5	250.0				75.0			###	32.1						173.8						
Holand-Netherlands																	21.4					
Portugal	141.1				155.6	107.1									166.7				236.7			
Yugoslavia																						42.1
South Korea																		870.0				
Honduras			945.0							151.7												
Cuba	31.8			19.0					501.5												28.9	
France														450.0			394.8				229.0	
Cambodia																125.0				750.0		
Dominican-Republic	146.0				116.7		75.0		375.0							92.7	38.1				35.1	
Laos		350.0											500.0			116.6					71.4	
Guatemala										136.2		25.8			121.8	47.5	39.8					
Columbia			175.0	80.3		79.0										46.6						
Ireland									500.0						100.0		533.3					
Trinadad&Tobago				175.0		333.3	200.0	89.3					920.0	66.3	63.8	466.7			250.0		453.0	243.8
Puerto-Rico	37.5	40.0	110.1	420.7		80.7	43.5	250.0	250.0				66.7		54.2	122.3	48.3	142.9	33.6	163.8	23.9	87.8
Ecuador			300.0	265.6	515.0	206.7	175.0	68.8	250.0		100.0			107.2	205.6		128.1	333.3	212.5		41.9	109.1
Peru	166.7	134.5			106.3	47.0	124.2	450.0	69.7					225.0		215.4	76.2	20.0	32.0	86.4	127.3	699.6
Nicaragua		74.1	178.3	65.6	140.0			47.6	83.3			160.0		159.5		340.0	76.5	85.7			81.0	
Mexico	122.2	159.1		89.7	155.2	67.6	40.3		75.0	59.9		17.1	46.5	34.5	95.0	61.9	59.8	121.7	82.1	140.3	52.9	89.6
El-Salvador				120.0	81.0	344.0	400.0			79.5			950.8	55.6	36.1	184.7	19.4		365.6	36.9	20.7	46.1
United-States	214.4	76.0	142.1	141.7	75.4	71.1	96.0	84.3	153.4				117.9	37.8	130.6	165.4	146.9	199.9	84.4	157.0	99.3	92.6

(b)

Figure 3: (a) Initial and (b) organized data representations of the *Census-Income*'s data cube.

According to the test of the Equation (3), for each $t \in \{1, \ldots, 5\}$, we select from D_t the set of characteristic attributes for the three selected factorial axes. These characteristic attributes give the best semantic interpretation of factorial axes 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 correspond to grayed rows in Table 1.

In the same way, we apply the test of the Equation (3) on the other dimensions of the cube. In the representation of the Figure 3(b), we clearly see that the regions of facts corresponding to characteristic attributes of the dimensions D_2 and D_5 seem to be more interesting and denser than other regions of the data space representation. These regions contains relevant information and reflect interesting association between facts. For instance, 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, *Asians* people from India, Iran, Japan and China are rather specialized in commercial jobs and trades.

7. EXPERIMENTAL RESULTS

We have realized some experiments on the *Census-Income* data cube presented in section 6. The aim of these experiments is to appreciate the efficiency of our approach by measuring the homogeneity gain realized by our MCA-based organization on data representations with different sparsity ratios. To vary sparsity we proceeded by a random sampling on the initial dataset of the 199523 facts from the considered cube.

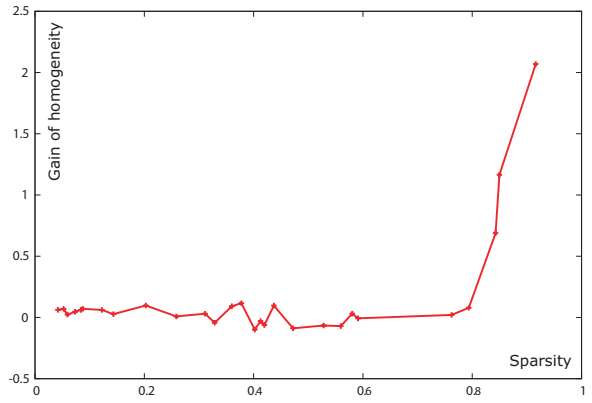

Figure 4: Evolution of the homogeneity gain according to sparsity.

According to Figure 4, 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 the data representation. For dense data cubes, the employment of our method is not always significant. This is naturally due to the property of the homogeneity criterion which closely depends on 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 Figure 4, we clearly see that curve is rapidly increasing to high positive values of gain when sparsity is greater than 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.

8. CONCLUSION AND FUTURE WORK

In this paper, we introduced a MCA-based method to enhance the representation of large and sparse data cubes. This method aims at providing an assistance to the OLAP user and helps him 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 an appropriate representation of OLAP facts. This representation provides better property for data visualization since it gathers full cells expressing interesting relationships of data. We also identify relevant regions of facts in this data representation by detecting characteristic attributes of factorial axes. 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 a concept of geometric neighborhood of cells. It also uses a similarity metric between cells. 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. We consider the problem of optimizing complexity of our approach. We also try to involve our approach in order to take into account the issue of data updates. Finally, we project to implement this approach under a Web environment that offers an interesting on line aspect and an interesting user interaction context.

9. REFERENCES

[1] 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.

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

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

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

[5] E. Malinvaud. Data Analysis in Applied Socio-Economic Statistics with Special Consideration of Correspondence Analysis. In *Marketing Science Conference*, Jouy en Josas, France, 1987.