New Data Warehouse Designing Approach Based on Principal Component Analysis

Wafa Tebourski,Wahiba Karaa, Wahiba Karaa
ISG, ENSI
Tunisia
wafatebourskiisg@yahoo.com, wahiba@time.tn, henda@ensi.tn

ABSTRACT: *Decision-making has become a strategic need for any business. Indeed, is among the capital business priorities. The establishment of decision information systems it facilitate the data exploitation and analysis. We distinguish data warehouses as the core system of business intelligence to ensure the structuring and analysis of multidimensional data. Consequently, the design of data warehouses has become a major problem, leading to the development of appropriate approaches to implement data warehouses. In this paper, we propose an approach to design and to construct data warehouses based on a descriptive statistics technique for the analysis of multidimensional data in the Principal Components Analysis (PCA). The findings of this article appear in two main areas: (i) a conceptual model data warehouse, (ii) an algorithm for the determination of measures and dimensions. A case study is used to validate our proposal.*

Keywords: Design of Data Warehouse, Mixed Approaches, Principal Components Analysis (PCA), Statistical Analysis, Correlation

Received: 29 May 2014, Revised 2 July 2014, Accepted 7 July 2014

1.Introduction

Decision Support Systems (DSS) deal with information from different sources in one place, consistent and familiar to the user. They combine and standardize databases, allowing analysis and decision making. Among the decision support systems, data warehouse systems are possibly the most used in the world wide. A data warehouse is "a *collection of data, integrated, non volatile and storied for decision making*" [16]. Since the 90s, data warehouses have become critical components of business intelligence. They have been successfully implemented in various sectors such as transportation, telecommunications, distribution, trade, medicine, financial services, insurance ... etc.

Data warehouses provide a broad vision of the company, an integration of different databases, a better organization and acquisition of the data, where the construction of a data warehouse is a daunting task especially effective warehouse schema designation. In fact, the search for a method of modeling data warehouses has become track booming. Several approaches for data warehouses design are proposed. They are classified into three categories: (i) approaches directed by the sources bottom-up), (ii) approaches directed by the needs (top-down) and (iii) mixed approaches. It should be noted that the bottom-up-approaches suffer from some limitations such as the inability of the decision maker to intervene on a conceptual level. The results may be schemas that do not satisfy their needs. In this approach, we find the generation of irrelevant schemas to make decisions. As for the top-down approaches, they require greater expertise of the designer in the modeling field. More generated models may not be satisfactory because the available data sources are heterogeneous, complex, and poorly

structured; which makes ETL process more difficult to achieve. In opposition, mixed approaches that support the needs and sources with the same level of importance provide better conceptual results.

Guided by mixed approaches, to benefit from their advantages, we introduce, in this paper, a new approach to data warehouse design based on principal component analysis, called DWDARPA. The originality of our approach is the statistical basis for our proposal because the principal component analysis is a descriptive method to summarize the original variables in a multidimensional reduced number of factors as their linear combinations. Indeed, this analysis gives several advantages namely, performance, flexibility and mathematical simplicity at its implementation.

The rest of the paper is organized as follows: In Section 2, we present a state of the art on the different approaches to design data warehouses. In Section 3, we introduce our approach. In Section 4, we present a case study on medical data dealing with cervical cancer to illustrate the proposed model. Finally, we conclude and indicate the future works in Section 5.

2. Related Work

Many researchers have focused on the design of data warehouse schemas, since it is a complex task. Our goal in this section is to present a comparative study between these different approaches, based on several criteria.

2.1 Sources Based Approaches

Sources based approaches are those which extirpate their data from dissimilar sources, which are then kept in a storage space accessible to all decision makers. The data warehouse design relies on an elaborate analysis of data models, mainly, the entityrelationship model (E/R). These approaches facilitate the process of ETL (Extraction Transformation-Load) as each entity and relation in the source model will be presented by multidimensional concepts.

We can cite different works conducted in this regard such as [3], [7], [8], [13], [14] and [15].

2.2 Requirements Based Approaches

Requirements based approaches are those which show the stages of the conceptual schemas requirement specification and derivation. Such approaches try to limit the decisional information system failure risk. Therefore, many works have been conducted in order to create requirementsbased data warehouses such as [5], [6], [9], [10], [11] and [12].

2.3 Mixed Approaches

Mixed approaches are those which include both bottom-up and top-down approaches in order to take advantage of their benefits. We can cite some studies in this respect, namely [1], [2] and [10].

Figure 1 shows a comparison between the different approaches of data warehouses design based on several criteria [17].

We can conclude that sources based approaches are useful if the data source schema is simple and available. In this class of approaches, there is still the risk of creating schemas that do not entirely meet the decision maker's needs. In contrast, the requirements based approaches focus on the requirements specification which is frequently variable and poorly expressed. Thus, the design of data warehouses cannot be exclusively based on data sources or requirements. Indeed, we find that both ascending and descending approaches are complementary and can be mixed together for better results. This is the objective of the third approach called hybrid approaches.

3. New Approach for Data Warehouse Design Based on Principal Component Analysis

We present in this section our functional architecture. Then, we present our process.

3.1 Functional Architecture

Figure 2 summarizes the different steps of our approach. Initially, the user declares its analytical objectives (facts). Starting with a set of entries such as the database and the external sources, the first step is to perform preprocessing on qualitative data that are stored in tables. The approach tests the similarity between the data based on the correlation between variables to group data into factors. These latter are translated into two types: (i) measures' factors include the heterogeneous data and (ii) dimensions' factors that support homogeneous data. Thereafter, our model studies the correlation between the obtained facts and factors. The last step ensures the schema generation of data warehouse that will be validated by the expert.

Approaches / Criteria		Approaches directed by sources			Approaches directed by requirements			Mixed approaches		
		Golfarelli	Hüsemann	Romero	Kimball	Cabibbo	Mazón	Bonifati	Nabli	Giorgini
	Logic schema	.	*	*	.	.	*	*	.	*
	Conceptual schema	.	.	*	*	.	.	*	.	*
Goals	Data Units	*	*	.	*	*	*	*		.
	Data Warehouses	*	*	.	*	*	*	*	.	*
Modeling	informal	*	*	*	.	.	*	*	*	*
	formal	.	.	.	*	*	.	.	.	.
Type of data sources	relational schema-	.	.	.	*	*			. .	
	ontology									
	ER diagram	.	. .	.		. DFM			. .	
Conceptual representation	UML	.	.		adopting	adopting				
	extension	.	.	*	.			.	.	
	others	*	*	.	.	.	.		.	
Method used for Requirement specification		.	.	.	.	Technique	TROPOS	GQM	.	TROPOS

Figure 1. Comparison between the different approaches to data warehouse design [17]

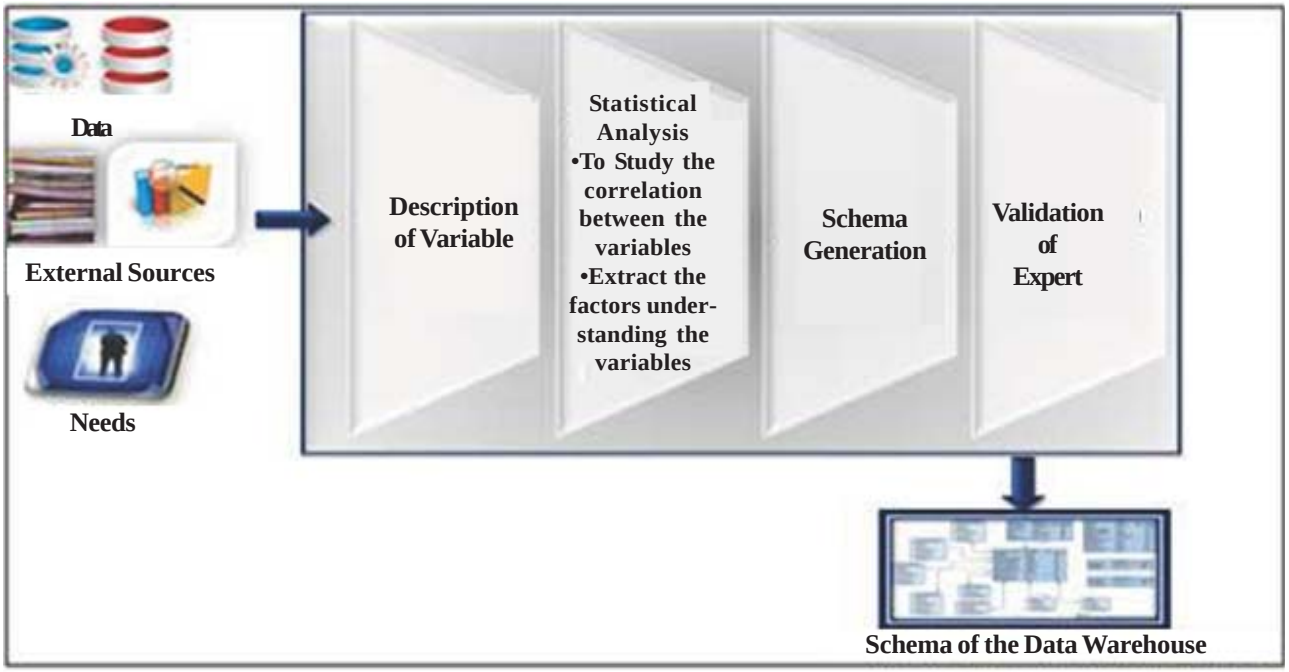

Figure 2. Functional architecture of our approach

3.2 Process

The process driving our DWDARPA approach has four steps that will be presented in the following subsections.

3.2.1 Description of Variables

As part of the description of the studied variables, we aim to transform the qualitative data into quantitative ones based on one of the following matrices: (i) contingency matrix: to cross two unimodal variables. If the co-occurrence measure applies to both unimodal variables, then we talk about a contingency measure, (ii) co-occurrence matrix: taking several representations depending on the purpose of the analysis. It is used to perform quantitative relational analysis, and (iii) presenceabsence matrix: crossing two variables. It records the existence of at least one individual.

3.2.2 Statistical Analysis

Our statistical analysis is reflected in the reduction of the data organized into a set of factors by minimizing the residual variance (intra-items variance) and maximizing the cumulative variance (inter-variable variance). This phase consists of two steps:

• In this step, we study the correlation between the variables using the correlation or the covariance matrixes to search for synthetic variables. Indeed, we use the correlation matrix when the variables are measured on different scales, while we use the covariance matrix when applying factor analysis to multiple groups with different variances for each variable. The data matrix must contain sufficient correlations to justify the link between the variables.

• In this step, we determine the number of factors to extract several criteria that are often chosen on the part of the variance of each item. Each factor chosen can explain:

(i) Rule Kaiser- Gutman which is: an eigen-value representing the amount of information captured by a factor. The latter, having an intrinsic value less than 1, represents less information than a simple item

(ii) Test Elbow: This test is fundamentally based on the factors' eigenvalues, but in a relative context and not in an absolute one. When, the amount of information is low or zero between two factors, we can estimate that the last factor does not contain sufficient information to extract the factors underlying the variables to be retained;

(iii) Percentage of variance shows the cumulative variance percentage extracted by the successive factors. The aim is to ensure that a factor explains a significant amount of variance. It is advisable to stop extracting factors when 60% of the explained variance is extracted [4].

3.3 Schema Generation

All components extracted from the candidate data warehouse schema that contain the fact summarizing the subject of analysis will include the dimensions that form the axes of the topic analysis from several perspectives.

3.4 Validation of Expert

Our approach is iterative and incremental. Indeed, the expert can validate the generated schema or check and restart another iteration of the process to achieve more satisfactory results.

4. Algorithm

Our new Data Warehouse Designing AppRoach based on Principal component Analysis, called DWDARPA, receives as input all the data set. It outputs the factors summarizing the most correlated variables, from which the data warehouse schema will be generated.

The notations used are shown in Table 1 and the pseudo- code of the algorithm is illustrated in the following sub section. In fact, DWDARPA, an iterative process, works in four stages: The first stage summarizes the data for the most explanatory variables and studies the correlation between these variables by calculating the correlation matrix. On the second step, we procede to the extraction of factors based on the correlated variables. At this level, we calculate the total variance reflecting the degree of information, that is the factor over all variables are included. Once the number of factors is determined, the expert can intervene to identify the obtained components. Finally, the data warehouse schema is generated and can be validated by the domain expert.

Notation	Description
Xi…n	All variables ranging from feature i to feature *n*
Cj	component *j*
C	All components
nC	Number of components
Corr(*k*, l)	Correlation between the two variables *k* and l
Mcorr	correlation matrix
Vt	Total variance
Var(*m*, *o*)	Cumulative variance of *m* and *o*

Table 1. List of Used Notations

Data : *Xi…n*
Result : *C* : components
Begin:
// Study the correlation between variables
For (*k* = 1 ; *k* < *n* ; *k* ++) **do**
For (*l* = 1 ; *l* < *n* –1 ; *l* ++) **do**
Mcorr = Calculate the correlation matrix
corr (*k*, l) = *calcul_correlation* (*k*, l)
Store corr (*k*, l) in *Mcorr*

// Retrieve the factors underlying variables
For (*m* =1 ; *m* < *n* ; *m* ++) **do**
Vt = Calculate the cumulative variance
Var (*m*, *o*) = *calcul_variance*
Store Var (*m*, o) in *Vt*

// Identify *C* from *Vt*
For (*i* = 1 ; *i* < *n* ; *i* ++) do
For (*j* = 1 ; *j* < *n* ; *j* ++) do
If *Vt* (*i*, *Cj*) is maximum then
Affect *i* in *Cj*
Return *C*
END

Algorithm: DWDARPA: New **D**ata **W**arehouse **D**esigning **A**pp**r**oach Based on **P**rincipal Component **A**nalysis.

5. Case Study

Our case study concerns the health of patients with breast cancer conducted by the University of Wisconsin Hospitals[1].

Sample code number	Clump thickness	Uniformity of the cell shape	Marginal adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	2
1002495	5	4	4	5	7	10	3	2	2
1015425	3	1	1	1	2	2	3	1	2
1016277	6	8	8	1	3	4	3	7	2
1017023	4	1	1	3	2	1	3	1	2
1017122	8	10	10	8	7	10	9	7	1
1018099	1	1	1	1	2	10	3	1	2
1018561	2	1	2	1	2	1	3	1	2
1033078	2	1	1	1	2	1	1	1	2
1033078	4	2	1	1	2	1	2	1	2
1035283	1	1	1	1	1	1	3	1	2
1036172	2	1	1	1	2	1	1	1	2
1041801	5	3	3	3	2	3	4	4	4
1043999	1	1	1	1	2	3	3	1	2
1044572	8	7	5	10	7	9	5	5	4
1047630	7	4	6	4	6	1	4	3	1
1048672	4	1	1	1	2	1	2	1	2
1049815	4	1	1	1	2	1	3	1	2
1050670	10	7	7	6	4	10	4	1	4
1050718	6	1	1	1	2	1	3	1	1
1054590	7	3	2	10	5	10	5	4	2
1054593	10	5	5	3	6	7	7	10	4
1056784	1	1	1	1	2	1	2	1	2
1057013	3	4	5	1	2	1	7	3	4
1059552	8	1	1	1	2	1	3	1	2
1065726	5	2	3	4	2	7	3	6	4
1066373	3	2	1	1	1	1	2	1	2
1066979	5	1	1	1	2	1	2	1	2

Table 2. Part of Our Dataset

5.1 Description of Variables

We investigate the health of 699 patients with breast cancer in 1991 presented by the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

The variables of our data set are, as shown by table 2:

- **Sample Code Number :** Patient code;
- **Clump Thickness:** Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer;
- **Uniformity of Cell Size/Shape:** Cancer cells tend to vary in size and shape. That is why these parameters are valuable in

[1]http://archive.ics.uci.edu/ml/datasets/Breast+Cance r+Wisconsin+(Original)

determining whether the cells are cancerous or not;

• **Marginal Adhesion:** Normal cells tend to stick together. Cancer cells tend to loose this ability;

• **Single Epithelial Cell Size:** Is related to the uniformity mentioned above. Epithélial cells that are significantly enlarged may be a malignant cell;

• **Bare Nuclei:** This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors;

• **Bland Chromatin:** Describes a uniform "*texture*" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser ;

• **Normal Nucleoli:** Nucleoli are small structures seen in the nucleus. Normally the cells nucleolus is usually very small. In cancer cells, the nucleoli become more prominent;

• **Mitoses:** The process in cell division by which the nucleus divides, typically consisting of four stages, prophase, metaphase, anaphase, and telophase;

• **Class:** 2 for benign and 4 for malignant.

5.2 Statistical Analysis

We in the following subsections present our conducted statistical analysis.

5.2.1 Study of the Correlation Between Variables

The purpose of this step is to summarize the data by forming a smaller number of more variables and more correlations. To do this, we use the correlation matrix that contains all the correlations betweenvariables (table 3). The Clump Thickness variable is positively correlated with the variables Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size and Mitoses, while Bare Nuclei variables Bland Chromatin are correlated with the variable Normal Nucleoli. This correlation between the variables will determine the set of variables that will compose the set of factors or components. All the correlated variables will be grouped into separate factors.

5.2.2 Extract Factors Using Variables: Total Explained Variance

The total variance gives us an idea about the level of information represented by each component or each factor. As shown in Table 4, we spent eleven input variables that were returned by four components. The first component accounts for 47.7% of the total information of all variables, the second includes 62.6%, the third 75.3% and the fourth includes 82.7%. Usually, we

	Sample code number	Chimp thickness	Uniformity Cell Size	Uniformity Cell Shape	Marginal Adhesion	Single Epethelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
n sample code number	1,000	-,055	-,042	-,042	-,065	-,046	-,099	-,050	-,052	-,035	-,080
Chimp thickness	-,055	1,000	,645	,645	,686	,522	,490	,458	,336	,550	,316
Uniformity Cell Size	-,042	,645	1,000	,907	,706	,752	,492	,356	,423	,539	,318
Uniformity Cell Shape	-,042	,655	,907	1,000	,683	,720	,312	,336	,319	,539	,319
Marginal Adhesion	-,065	-,686	,706	,683	1,000	,600	,364	,267	,203	,518	,497
Single Epethelial Cell Size	-,046	,522	,752	,720	,600	1,000	,384	,316	,229	,679	,483
Bare Nuclei	-,099	,490	,492	,312	,364	,384	1,000	,674	,587	,338	,316
Bland Chromatin	-,060	,458	,356	,336	,267	,316	,674	1,000	,666	,344	,357
Normal Nucleoli	-,052	,336	,423	,319	,203	,229	,587	,666	1,000	,428	,412
Mitoses	-,035	,550	,559	,539	,518	,679	,338	,344	,428	1,000	,423
Class	-,080	,316	,318	,319	,497	,483	,316	,357	,412	,423	,1000

Table 3. Correlation Matrix of Our Dataset

choose the variables that have a total superior to 1. In our case, we consider four components presenting 82.7% of the total information. The expert can intervene at this stage to name the generated components.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	**5, 724**	**47, 688**	**47, 688**	**6, 724**	**47, 688**	**47, 688**
2	**2, 1001**	**14, 894**	**62, 582**	**21, 001**	**14, 894**	**62, 582**
3	**1, 792**	**12, 709**	**75, 291**	**1, 792**	**12, 709**	**75, 291**
4	**1, 047**	**7, 425**	**82, 716**	**1, 047**	**7, 425**	**82, 716**
5	**0, 759**	**5, 383**	**88, 099**			
6	**0, 579**	**4, 106**	**92, 206**			
7	**0, 317**	**2, 248**	**94, 454**			
8	**0, 298**	**2, 113**	**96, 567**			
9	**0, 262**	**1, 858**	**98, 426**			
10	**0, 133**	**0, 943**	**99, 369**			
11	**0, 089**	**0, 631**	**1, 00, 000**			

Table 4. Total Variance

	Component			
	1	2	3	4
Sample code number	-, 083	, 992	-, 073	, 008
Climp Thickness	, 742	-, 007	-, 134	, 609
Uniformity of Cell Size	, 922	, 048	-, 020	-, 043
Uniformity of Cell Size	, 915	, 043	-, 054	-, 003
Marginal Adhesion	, 802	-, 009	-, 003	-, 309
Single Epithelial Cell Size	, 807	, 047	, 173	-, 108
Bare Nuclei	, 315	-, 073	-, 217	, 824
Bland Chromatin	, 469	„ 001	-, 172	, 839
Normal Nucleoli	, 223	, 028	,043	, 812
Mitoses	, 844	, 059	, 200	, 115
Class	, 534	-, 022	, 920	, 100

Table 5. Component Matrix After Rotation

5.2.3 Component Matrix After Rotation

Table 5 shows the correlation between variables and the four components identified in the previous step. Examination of the correlations between the original variables and the principal components allows interpreting them and the corresponding principal axes. Axis 1, named "*CELL*", includes variables Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size and Mitoses. In fact, these variables are most correlated with the fourth component Bare Nuclei mainly formed by the variable Sample code number. They are intended to illustrate the various states of patients affected by breast cancer. Axis 2, named "*CLASS*", represents the class of variable. This variable shows the two class types of breast cancer that the patient may have, namely: "*benign*" coded 2 and "*malignant*" coded "4". Axis 3, called "*NUCLEI*", includes the, Bland Chromatin and Normal Nucleoli variables that are correlated with the first component. Axis 4, named "*PATIENT AFFECTED*", includes the variable Sample code number.

5.2.4 Generation Schema of the Data Warehouse

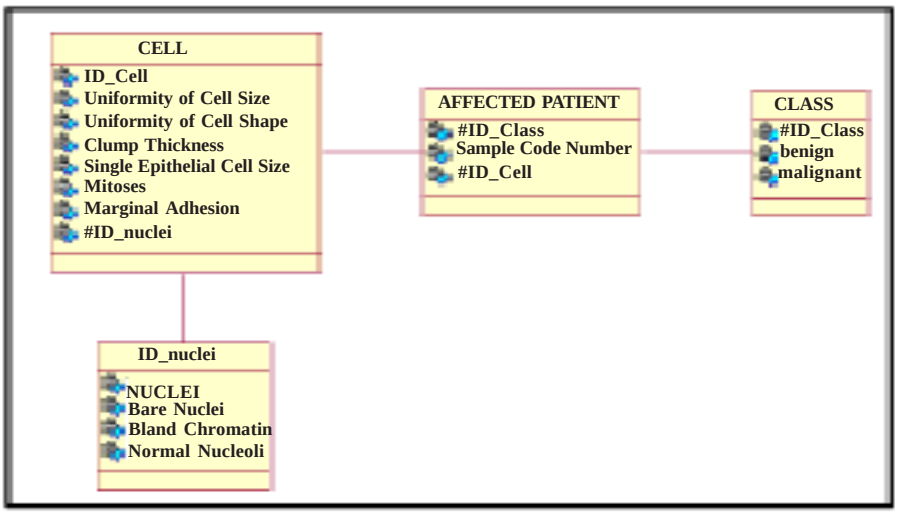

Figure 3. Snowflake schema data warehouse on the cancer of breast

The logical schema of obtained data warehouse is presented in figure 3. Indeed, it is composed of a fact related to "*PATIENT AFFECTED*". Among, the dimensions considered is the "*CELL*" dimension, each cell is described by their Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size and Mitoses. The dimension "*CLASS*" is described by the different class of breast cancer. The dimension "*NUCLEI*" describes the state of the cancerous cell via the attributes Bare Nuclei, Bland Chromatin and Normal Nucleoli.

6. Conclusion

In this paper, we proposed a new approach for data warehouses building based on the fundamentals of descriptive statistics for the analysis of multidimensional data PCA. To do so, we introduced our functional architecture that is driven by an original process. In addition, a new algorithm is presented. To illustrate our proposal, a case study is detailed. Other prospects for future work mainly concerned the following tracks: (i) the use of other statistical techniques such as indices and (ii) the support of the process using domain ontology to minimize recurrent intervention of the expert.

References

[1] Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S. (2001). Designing data marts for data warehouses. ACM Trans. Softw.Eng. Methodol.

[2] Nabli, A., Feki, J., Gargouri, F. (2005). Automatic Construction of Multidimensional Schema from OLAP Requirements. Arab International Conference on Computer Systems and Applications (AICCSA'05).

[3] Hüsemann, B., Lechtenbörger, J., Vossen, G. (2000). Conceptual Data Warehouse Design. Design and Management of Data Warehouses.

[4] Jr.Hair, J. F., Black, W. C., Babin, Anderson, R. E., Tatham, R. L. (2006). Multivariate Data Analysis, 5e édition , Ed. Pearson-Prentice Hall, New Jersey.

[5] Mazón, J., Trujillo, J., Serrano, M., Piattini, M. (2005).Designing data warehouses: from business requirement analysis to multidimensional modeling, in: Cox, K., Dubois, E., Pigneur, Y., Bleistein, S. J., Verner, J., Davis, A. M., Wieringa, R. (Eds.), REBNITA Requirements Engineering for Business Needs and IT Alignment, University of New South Wales Press.

[6] Cabibbo, L.,Torlone, R.(1998). A Logical Approach to Multidimensional Databases. In VIth International Conference on Extending Database Technology (EDBT 98).

[7] Golfarelli, M., Maio, D., Rizzi, S. (1998). The dimensional fact model: conceptual model for data warehouses. *International Journal of Cooperative Information Systems* 7.

[8] Romero, O., Abelló, A. (2007). Automating Multidimensional Design from Ontologies. DOLAP'07.

[9] Giorgini, P., Rizzi, S., Garzetti, M. (2005). Goaloriented Requirement Analysis for Data Warehouse Design. *In*: Proc. of 8th Int.Workshop on Data Warehousing and OLAP.

[10] Giorgini, P., Rizzi, S., Garzetti, M. (2008). A Goal- Oriented Approach to Requirement Analysis in Data Warehouses. In Decision Support Systems (DSS) jounal.

[11] Vassiliadis, P., Simitsis, A., Skiadopoulos, S. (2002). Conceptual modeling for etl processes, in Theodoratos.

[12] Kimball, R. (1996). The Data Warehouse Toolkit, John Wiley and Sons, Inc., New York.

[13] Luján-Mora, S., Trujillo, J., Song, I. Y. (2002). Extending the UML for multidimensional modeling. *In*: Proceedings of the International Conference on the Unified Modeling Language. Dresden, Germany.

[14] Luján -Mora, S., Trujillo, J., Song, I. Y. (2006). A UML Profile for Multidimensional Modeling in Data Warehouse. Data and Knowledge Engineering.

[15] Rizzi, S. (2007).Conceptual Modeling Solutions for the Data Warehouse. Data Warehouses and OLAP: Concepts, Architectures and Solutions, edited by Wrembel, R., Koncilia, C.

[16] Inmon, W. H. (1996). Building the Data Warehouse. John Wiley & Sons.

[17] Tebourski, W., Karrâa, W., Ben Ghezala, H. (2013). Semi automatic Data Warehouse Design methodologies: a survey.IJCSI *International Journal of Computer Science Issues.*