Modeling Heterogeneous Dimensions in Multi-Dimensional Databases

Ali A. Safaei
Dep. of Computer Engineering,
Iran Uni. of Science and Technology
Tehran, Iran.

Reza Pirshayan
Dep. of Computer Engineering,
Iran Uni. of Science and Technology
Tehran, Iran.

Mostafa S. Haghjoo
Dep. of Computer Engineering,
Iran Uni. of Science and Technology
Tehran, Iran.

***Abstract*-The need for analyzing huge amount of data in organizations as well as effects of analyses results on its decision making and strategic planning in growing. Multi-dimensional databases issue this problem. Multi-dimensional conceptual model strongly affects multi-dimensional database development. Many of the existing multi-dimensional models are based on traditional relational model while some others aim to organize data using a multi-dimensional perspective. These models do not support heterogeneous dimensions and this is a major deficit, causing limitations in their applications. In this paper, a conceptual model is proposed to be able to model heterogeneous dimensions. The concept of summarizable hierarchy proposed and is used to model heterogeneous dimensions in multi-dimensional databases. The proposed model is shown to be more suitable for modeling real-world applications.**

Keywords: multi-dimensional databases, OLAP, summarizability, heterogeneous dimensions.

I. INTRODUCTION

Traditional models such as relational or ER are profitable for transactional processing applications (several concurrent transactions each with a relatively few amount of processing) but not for analytical processing applications (few transactions each with a large volume of processing). Multi-dimensional models are introduced for these applications.

Multi-dimensional models categorize data as facts, measures and dimensions. As an example, in a small business, products are sold with specified time, amount and price. “Selling” could be considered as the fact, “amount” and “price” as measures and “products type” and “time” as dimensions.

Dimension is the most essential element of a multi-dimensional model. Existing multi-dimensional models are unable to support heterogeneous dimensions and this is a major deficit for modeling real-world applications. In this paper, a multi-dimensional model is proposed that supports heterogeneous dimensions.

The rest of the paper is organized as follow: related work is summarized in section 2. In section 3, hierarchy of dimension and its categorization and the concept of summarizability are presented. The proposed multi-dimensional model is proposed in section 4 and evaluated in section 5. Finally, we conclude in section 6.

II. RELATED WORK

Despite traditional conceptual models, multi-dimensional models are introduced to be employed in analytical applications. Most of the multi-dimensional models concentrate on query’s properties while just a few on semantics.

Agrawal, Gupta and Sarawagi proposed a data model and some operators to specify basic concepts of multi-dimensional databases [1]. This model has symmetric behavior for all of the dimensions and measures. It supports multiple hierarchy-levels for each dimension. Also, the operators provide the ability to express multi-dimensional queries.

Gyssens and Lakshmanan proposed a multi-dimensional model which can be employed in OLAP applications [2]. It provides a transparent definition to distinguish between structural and conceptual aspects. This enriches the model for specifying data manipulating language.

In [3], Cabbio and Torlone proposed a logical model for OLAP systems which opposed to the other multi-dimensional models, is implementation-independent. They introduced methods for designing multi-dimensional database schema and transforming a multi-dimensional database to the corresponding multi-dimensional arrays and tables in relational model.

III. SUMMARIZABILITY OF DIMENSIONS

In a multi-dimensional database, vast amount of semantically and potentially relevant data are managed. These data are organized, retrieved and analyzed in different perspective, known as dimension [8]. Efficient categorization of dimensions and hierarchy of dimensions is introduced in [4-7]. Deduction of summarized level from dimensional hierarchy-levels is performed using split constraints in a manner proposed in [4-7].

A. *Dimension*

- *Hierarchy schema*

  Hierarchy schema is a tuple as G=(L,↑, A, σ) in which:
  L is set of levels that has a special level ALL
  ↑ is a binary relation on L such that (L,↑) compose a DAG with the node ALL as the root. Also, ↑* is transitive closure of ↑.
  A is set of attributes
  σ: L→2A is mapping of each level to a set of attributes.

Bottom levels of G are in form of $L_{Bottom} = \{l \in L | \neg\exists l' \in L: l' \uparrow l\}$. Also, $\gamma(l_a, l_b)$ indicates set of paths between levels l_a and l_b in G.

- *Dimension instance*

  A dimension instance is a tuple as τ=(G,ε,<,T) in which:

  G: dimension schema

  ε: set of elements set in E; ε_l indicates elements set of level l ∈ L and ε_d is union of elements sets of dimension d.

  < is relation between elements of DAG and the root ALL. Also, << is its transitive closure.

  T: set of relations that for each level l ∈ L it contains a relation T_l with attributes $\sigma(l) \cup \{l\}$. The following conditions will be hold:

  Element set of level l (*i.e.,* ε_l) is the active level of L in T_l.

  $\forall(e_a, e_b) \therefore (e_a \in e_{l_a}) \wedge (e_b \in e_{l_b}) \wedge ((e_a < e_b)) \Rightarrow l_a \uparrow l_b$

  and there exist elements as $e_1, e_2, \ldots, e_n \in e_d$ such that

  $e_a < e_1 < e_2 < \cdots < e_n < e_b$

$$\forall e \in e_d \therefore e \neq \text{ALL} \Rightarrow e \ll \text{ALL}$$

  The base level that is denoted as l_{base} is the union of bottom level elements sets.

Figure 1 shows an example of dimension schema (figure 1(a)) and dimension instance (figure 1(b)).

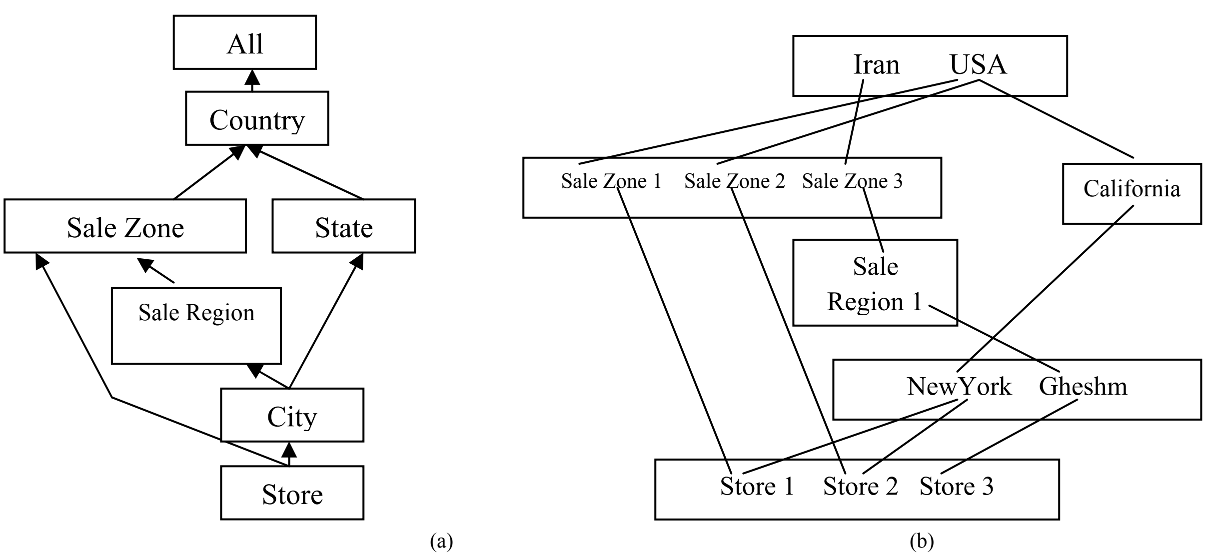


Figure1: an example of (a) dimension schema and (b) dimension instance

Definition 1: Rollup operator

Let h is a dimension instance. A direct rollup operator gives a relation with attributes l_1 and $l_2$from arguments of type levels (l_1 and l_2) of a dimension.

$$d\Gamma_{\ell_1}^{\ell_2} = \{(x_1, x_2) \mid x_1 \in \varepsilon_{\ell_1} \wedge x_2 \in \varepsilon_{\ell_2} \wedge x_1 < x_2\}$$

Accordingly, rollup operator is defined via employing transitivity:

$$\Gamma_{\ell_1}^{\ell_2} = \{(x_1, x_2) \mid x_1 \in \varepsilon_{\ell_1} \wedge x_2 \in \varepsilon_{\ell_2} \wedge x_1 << x_2\}$$

$$\Gamma_{\ell_{base}}^{\ell} = \{(x, y) \mid x \in \varepsilon_{\ell_{base}} \wedge y \in \varepsilon_{\ell} \wedge x << y\}$$

The following conditions are hold:

$if\ \neg(\ell_a \uparrow \ell_b)$ then $\mathrm{d}\Gamma_{\ell_a}^{\ell_b} = \varnothing$

$if\ \neg(\ell_a \uparrow^* \ell_b)$ then $\mathrm{d}\Gamma_{\ell_a}^{\ell_b} = \varnothing$

$if\ \gamma_{\ell_a, \ell_b} = \{\ell_a, \ell_b\}$ then $\Gamma_{\ell_1}^{\ell_b} = d\Gamma_{\ell_a}^{\ell_b}$

- *Dimension schema*

  Dimension schema is a tuple as ds=($G,\sum$) in which G is the dimension schema and $\sum \subseteq CL$ (is a subset of constraint languages).

Definition 2: Homogeneous dimension

Dimension d is homogeneous if for its dimension schema ds:

$\forall(l_1, l_2) \therefore l_1, l_2 \in L \Rightarrow \Gamma_{l_2}^{l_1}: \varepsilon_{l_1} \rightarrow \varepsilon_{l_2}$ is a total function.

Definition 3: Heterogeneous dimension

Dimension d is heterogeneous if it is NOT homogeneous [4].

Figure 2 illustrates dimension schema for example of figure 1.

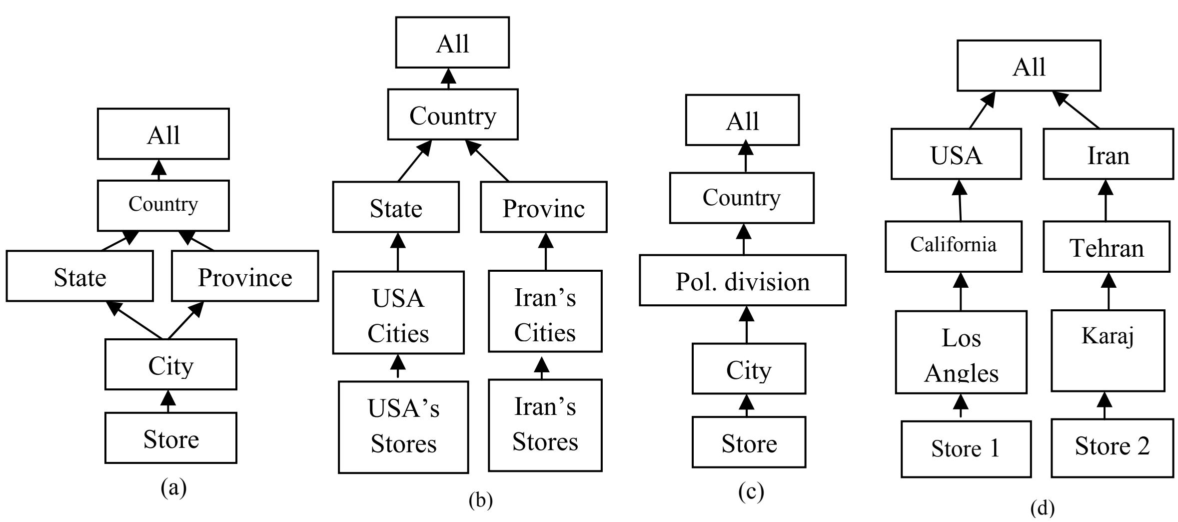


Figure 2: homogeneous and heterogeneous dimensions for example of figure 1

Dimension schema in figure *2(b)* is homogeneous but in figure *2(a)* is heterogeneous because in instance of figure *2(d)*, rollup function between city and state is partial function from *{NewYork, Toronto}* to *{NYstate}*.

Definition 4: Hierarchical dimension schema

Dimension schema *ds* is hierarchical if for each two levels $l_1, l_2 \in L_{ds}$ such that $l_1 \uparrow l_2$, for each dimension instance of *ds*, $l_1 < l_2$.

Accordingly, *ds* is strict hierarchical if $\forall l_1, l_2 \in L_{ds} \therefore l_1 \uparrow^* l_2 \Rightarrow \forall \tau_i \in \tau(l_1 < l_2)$

B. Summarizability

Summarizability specifies conditions in which a cube view in level l_a is correctly mapped to cube view in level $l_b$in a specific dimension [8]. The challenge is to investigate summarizability without investigating corresponding dimension instance. For example, in figure *1(b)*, considering dimension schema, we can find that it is summarizable and total sale value for each state can be computed from sale value of cities.

Definition 5: Summarizable

Level l in dimension instance *d* for distributive function *f* is summarizable if

$$cv(d, F, \ell, af(m)) = \prod_{\ell, af^c(m)} (\bigcup_{i \in 1..n} (\prod_{\ell, m} (\Gamma_{\ell_i}^{\ell} \infty\, cv(d, \ell_i, af(m))) .$$

In figure 1(b), "*country*" is summarizable from {*state, sale region*} but is not summarizable from {*state, sale zone*}.

Lemma 1: *level $l \in L$, L={ l_1, l_2,..., l_{1n} } in dimension instance d is summarizable iff*

$$\Gamma_{\ell_{base}}^{\ell} = \bigcup_{i \in 1..n} \prod_{\ell_{base}, \ell} (\Gamma_{\ell_{base}}^{\ell_i} \infty\, \Gamma_{\ell_i}^{\ell}) .$$

IV. THE PROPOSE MODEL

The homogeneous dimensions, correct conditions of summarizability can be investigated w.r.t. edges of the dimension hierarchy graph. But in heterogeneous dimensions, dimension hierarchy graph doesn't indicate summarizability relationship. So, we introduce hierarchy schema and summarizable hierarchy schema and show that we will be able to model heterogeneous dimensions as well as homogeneous dimensions.

A. Dimensional hierarch schema

Definition 6: Dimensional hierarchy schema

Dimensional hierarchy schema is a tuple as *h*=($L_h, <, R - UP_h, \Delta$) in which:

$$L_h \subseteq h$$

< is a partial order relation on L_h:

$$\forall l, l' \in L_h \therefore \exists x \in l, x' \in l' \wedge x < y \Rightarrow l < l'$$

$R - UP$(rollup function): for each two levels l_1, l_2 such that $l_1 < l_2$, rollup function $R - UP_{l_1}^{l_2}$maps elements from DOM(l_1) to an element in DOM(l_2) according to Δ .

Δ is set of rules for rolling up the elements of a level to other levels.

$O_1$is called to be rolled up to O_2*when* $R - UP_{l_1}^{l_2}(O_1) = O_2$.

Definition 7: summarizable dimensional hierarchy schema

Dimensional hierarchy schema *h*=($L_h, <, R - UP_h, \Delta$) is summarizable if Δ contains the following rules:

$$\Delta = \begin{cases} \forall \ell, \ell_i \in L_h, \ell \prec \ell_i, \forall x \in DOM(\ell), \exists y : \bigoplus_{i=1}^{n} R - UP_{\ell}^{\ell_i}(x) = y \wedge y \in DOM(\ell_i) \ (1 \le i \le n) \\ \forall y \in DOM(\ell), \ell \ne \ell_{base}, \exists \ell' \in L, x \in DOM(\ell') : R - UP_{\ell'}^{\ell}(y) = x \wedge \ell' \prec \ell \end{cases}$$

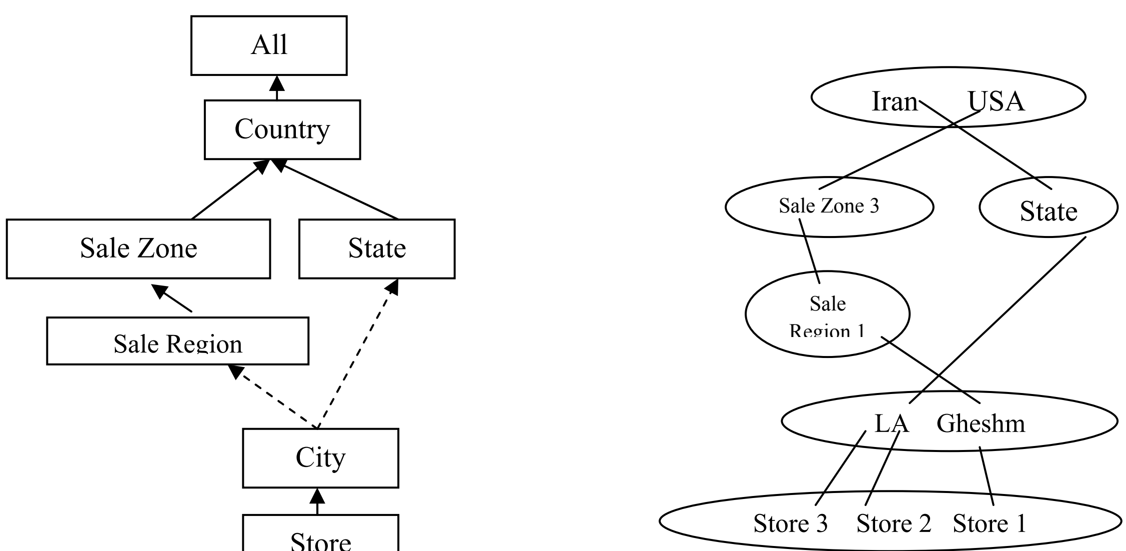
So, summarizable dimensional hierarchy schema is stated as $sh=(L_{sh}, \prec, R-UP_{sh})$. For example, this dimensional hierarchy schema (derived from figure 1(a)) is summarizable:

$$
\begin{aligned}
&L_{h1} = \{StoreId, City, Province, Country, All\}\\
&\prec = \{(StoreId \prec City), (City \prec Province), (Province \prec Country), (Country \prec All)\}\\
&\Delta :\\
&\forall x \in StoreId, \exists y \in City : R-UP_{StoreId}^{City}(x) = y\\
&\forall x \in City, \exists y \in Province : R-UP_{City}^{Province}(x) = y\\
&\forall x \in Province, \exists y \in Country : R-UP_{Province}^{Country}(x) = y\\
&\forall x \in Country : R-UP_{Country}^{All}(x) = all.
\end{aligned}
$$

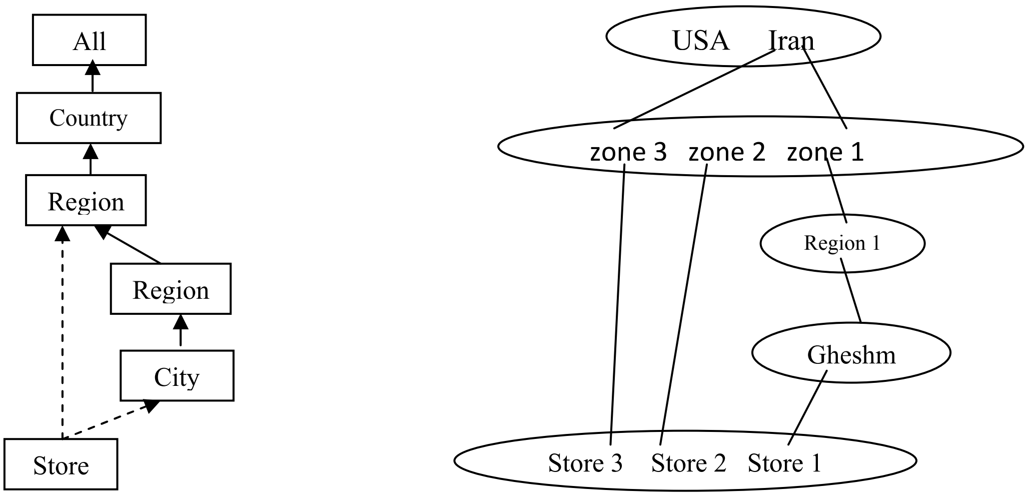
And the following (which is derived from figure 1(b)) is summarizable, too.

$$
\begin{aligned}
&L_{h2} = \{Store, City, State, Province, Country, All\}\\
&\prec = \{(Store < City), (City < State), (City, provience), (State, Country), (province, Country)\}\\
&\Delta :\\
&\forall x \in Store, \exists y \in City : R-UP_{Store}^{City}(x) = y\\
&\forall x \in City, \exists y : (y \in Province \wedge R-UP_{City}^{Province}(x) = y) \oplus (y \in State \wedge R-UP_{City}^{State}(x) = y)\\
&\forall x \in Province, \exists y \in Country : R-UP_{Province}^{Country}(x) = y\\
&\forall x \in State, \exists y \in Country : R-UP_{State}^{Country}(x) = y\\
&\forall x \in Country : R-UP_{Country}^{All}(x) = all.
\end{aligned}
$$

Figure 3 shows summarizable heterogeneous schema and instances of figure *2(a)*.



(a) (b)



(c) (*d*)

Figure 3: summarizable heterogeneous schema and instances of figure 2(a)

Dashed lines in figure 3 indicate that not all of the elements in the lower levels are rolled up to the upper levels. So, summarizability can be induced from summarizable hierarchy schema; each level is summarizable from levels which are as source node of an edge in DAG that is connected to that level.

$$\forall l_i, l_j \in L: l_i \text{ is summarizable from } l_j \Leftrightarrow \exists (l_i, l_j) \in G$$

With respect to definition of summarizable dimensional hierarchy schema, *Sh1 =(L,<, R-UP)* in which:

$$L = \{StoreId, City, SaleDistrict, SaleRegion, State, Country, All\}$$
$$\prec = \{(StoreId, City), (City, State), (City, SaleDistrict), (SaleDistrict, SaleRegion), (SaleRegion, Country), (Country, All)\}$$

is summarizable dimensional hierarchy schema of figure *3(a)* and *sh2* in which:

$$L = \{StoreId, City, SaleDistrict, SaleRegion, Country, All\}$$
$$\prec = \{(StoreId, City), (StoreId, SaleRegion), (City, SaleDistrict), (SaleDistrict, SaleRegion), (SaleRegion, Country), (Country, All)\}$$

is summarizable dimensional hierarchy schema of figure *3(c)*:

V. Evaluation

The most important metrics for evaluation of a general multi dimensional model are described below:

- *Implementation independency:* The model must be conceptual and independent of implementation details.
- *Isolation of syntax (structure) and semantic*: Structure that is used for representation of data must be independent from semantics of data.
- *Explicit hierarchy of dimensions support:* Ability of defining hierarchy in each dimension explicitly should be supported.
- *Multiple-hierarchies in dimensions support:* For example, it is desirable to have two rolling up paths: day→ month→ season→ year and day→ week→ year
- *Summarizability support:* Correctness of rolling up in dimensional hierarchy (*e.g.,* avoiding to compute a value more than once while rolling up).

- *Heterogeneous dimensions support*

According to these metrics, the proposed model is compared with the other general multi dimensional models as described in table 1.

TABLE 1: COMPARISON OF THE PROPOSED MODEL WITH THE OTHERS (*P* MEANS PARTIALLY SUPPORT)

Metric / Model	A	B	C	D	E	F
Li and Wang [10]	-	✓	*P*	✓	-	-
Gyssens and Lakshmanan [2]	-	✓	-	✓	-	-
Agrawal et. al. [1]	✓	-	✓	✓	-	-
Cabbio and Torlone[3]	✓	✓	✓	✓	-	-
Lehner et. al. [11]	✓	-	✓	-	*P*	-
Golfarelli et. al. [12,13]	✓	✓	✓	✓	-	-
Sapia et. al. [14]	✓	✓	✓	✓	-	-
Franconi and Sattler [15]	✓	✓	✓	✓	-	-
The proposed model	✓	✓	✓	✓	✓	✓

VI. CONCLUSION

Volume of data and diversity of analytical applications are growing rapidly and database paradigms such as multi-dimensional database are very beneficial. Traditional data models don’t satisfy multi-dimensional databases requirements. On the other hand, multi-dimensional models don’t support heterogeneous dimensions which are required to be modeled in many-real-world applications. In this paper, a multi-dimensional model is proposed that supports heterogeneous dimensions. Concept of summarizable hierarchy is defined which can be used to decompose heterogeneous dimensions into summarizable components. The proposed model support important metrics required for modeling and covers a vast range of real-world applications.

REFERENCES

[1] R. Agrawal, A. Gupta, S. Sarawagi : “Modeling Multidimensional Databases”, In Proceedings of 13th International Conference on Data Engineering, Birmingham, U.K, 1997, 232-243.

[2] M. Gyssens, L. Lakshmanan : “A Foundation for Multidimensional Databases”, International Conference on Very Large Data Bases(VLDB), Athens, Greece, August, 1997, 25-29.

[3] L. Cabbibo, R. Torlone : “A Logical Approach to multidimensional databases”, 6th International Conference on Extending Database Technology (EDBT 98), Springer-Verlag, 1998.

[4] Carlos A. Hurtado, Alberto O. Mendelzon : “Reasoning about Summarizability in Heterogeneous Multidimensional Hierarchies”, Proceedings of the 8th International Conference on Database Theory, Jan. 2001, 375-389.

[5] Carlos A. Hurtado, Alberto O. Mendelzone : “OLAP Dimension Constraints”, In Proceedings of the 21st ACM SIGMOND-SIGACT-SIGART Symposium on Principles of Database Systems, Madison –Wisconsin,USA,Jun. 2002, 169-179.

[6] Carlos A. Hurtado, C. Gutiérrez : “Equivalence of OLAP Dimension Schemas”, In Proceedings of 3rd International Symposium, Folks 2004, Wilhelminenburg Castle, Austria, Feb. 2004, 176-195.

[7] Carlos A. Hurtado, C. Gutiérrez , Alberto O. Mendelzone, “Capturing summarizability with integrity constraints in OLAP”, ACM Transactions on Database Systems, Volume 30 , Issue 3, 2005

[8] M. Rafanelli, A. Shoshani : "STORM: A Statistical object Representation Model", In Proceedings of the 5th SSDBM Conference, Charlotte, N.C.,USA, 1990.

[9] Kenan Technologies: "An Introduction to Multidimensional Databases", 1996. Technical Report.

[10] C. Li, X. S. Wnag, "A Data Model for Supporting On-Line Analytical Processing", Conference on Information and Knowledge Management, Nov. 1996.

[11] W. Lehner, J. Albrecht, H. Wedekind : "Normal Forms for Multidimensional Databases", In 10th SSDBM Conference, Italy, July 1998.

[12] M. Golfarelli, S. Rizzi : "A methodological Framework for Data Warehouse Design", First International Workshop on Data Warehousing and OLAP (DOLAP), ACM, Nov 1998.

[13] M. Golfarelli, D. Maio, S. Rizzi : "Conceptual Design of Data Warehouses from E/R Schemas.", International Conference on Systems Science, Hawaii, IEEE, Ene. 1998.

[14] C. Sapia, M. Blaschka, G. Hoeffling : "GraMMi: Using a Standard Repository Management System to Use a Graphical Modeling Tool" , Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8, Jan. 2000.

[15] E. Franconi, U. Sattler, "A Data Warehouse Conceptual Data Model for multidimensional Aggregation.", Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW 99), S. Gatziu, M. Jeusfeld, M. Staudt, Y. Vassiliou. Eds., Jun 1999.