A functional model for macro-databases

M. Rafanelli (*) and F.L. Ricci (^)

(*) Istituto di Analisi dei Sistemi ed Informatica, C.N.R., viale Manzoni 30, 00185 Roma, Italy
(^) Ist. di Studio per la Ricerca e la Docum. Scientifica, C.N.R., via De Lollis 12, 00185 Roma, Italy

Abstract *Recently there have been numerous proposals aimed at correcting the deficiency in existing database models to manipulate macro data (such as summary tables). The authors propose a new functional model, Mefisto, based on the definition of a new data structure, the "statistical entity", and on a set of operations capable of manipulating this data structure by operating at metadata level.*

1 Introduction

Statistical databases (SDBs) are widely used in applications, such as census data analysis, economic planning, health care organizations, etc. They are different from conventional DBs in the following ways [Shos 82], [ShWo 85]:

a) *the data structure* - most of the existing models for conventional DBs support merely *simple data structures*, as the "relations", whereas SDBs need to support *complex data structures* [BrNS 83].

b) *the data manipulation* - boolean operations or logical associations between data are not of prime importance to statisticians; in fact, the most common manipulation is related to the encoding of data, or to the reclassification of the descriptive data [Gosh 86].

There are two broad classes of SDBs, micro and macro SDBs [Wong 84]. The former (micro-DBs) refers to SDBs containing micro data, that is, records of individual entities or events (such as, mortality data of individual people or population census). The latter (macro-DBs) refers to SDBs containing macro data, often shown as summary tables, that result from the application of aggregate functions (for example, count, sum, or average) on data of micro-DBs (such as tables of "consumption of energy" or charts of "mortality by disease"). These two classes of SDBs are quite different: the more relevant difference is the existence of an intentional and an extensional level in the metadata of macro data.

In this paper we propose a statistical functional model for macro data (Mefisto) in which new operations are defined; these operations carry out the statistical analysis (which however turns out to be a subsequent phase, tied to the use of statistical packages). The Mefisto model presents the advantages of the flexibility, compactness, and the use of the operations, in that they are independent from the single SE and the single summary type, and of the simplicity of use for the statistical user (the *objects* described by the Mefisto model come close to the statistical user's way of thinking).

2. MODELING STATISTICAL MACRO DATA

We will call "statistical entity" (SE) any representation of data structures for summary data (relation, vector, summary table, time series, etc.) in statistical databases (SDBs). The elements that characterize a SE are:

a) a single *summary attribute* representing a property of the phenomenon described in the SE; its instances (summary values) are the numeric values inside the summary table. Its *summary type* depends on the particular aggregate function that generated it; for example, the aggregate function "percentage" produces, as a summary type, "rate".

b) a set of *category attributes* which have the role of characterizing the summary attribute. It is the intentional level of metadata.

c) *statistical entity variable domain*, that is, a set of values corresponding to every category attribute; such values are generally strings of alphanumeric characters. It is the extensional level of metadata.

The Cartesian product of all the statistical entity variable domains of the SE represents the *statistical entity space;* the summary values of the SE are all determined by the elements of this space.
In Fig.1 an example of SE is shown.
The summary attribute is "cancer around the world" and its summary type is "rate per 100,000" (the aggregative activity in this application is the function "count" applied to the above micro data set and the subsequent function "ratio", generating the result per 100,000). The set of category attributes is composed of "sex", "country", and "site" of the cancer; the statistical entity variable domain of sex is {M (= male), F (= female)}, the statistical entity variable domain of site is {oral, lung}, and the statistical entity variable domain of country is {Australia, Austria, Denmark}.

The activity of users of macro-DBs generally includes two

Cancer around the world		site			
		oral		lung	
rate per 100000		sex M	F	sex M	F
country	Australia	5,1	1,7	71,6	21,4
	Austria	5,3	0,6	65,9	9,7
	Denmark	4	2	70,6	18,7

Source of data: World Health Organ., Annual Statistics, 1984

Figure 1

orthogonal phases which are characterized by the following:

1) the manipulating of the descriptive part of the macro data, that is the metadata *(Statistical Entities Manipulating);*

2) the processing of the summary values *(Data Analysis).*

The former is characterized by the homogenization of the macro data which are often distributed among different data sources, and by verifying the semantic consistency and comparability among different levels of aggregation of the available data.
The latter processes the data by statistical-mathematical functions (such as regression, clustering etc.); in this last phase the statistical expertise is also important.

3 The Mefisto model

The Mefisto model is a logical model based on the functional approach, which represents the macro datum independently of its physical storage or its display form to the user. It is concerned with the logical management of the SE (statistical entity management); the user manipulates the descriptive elements of the SEs, changing their characteristics both at the level of category attributes (for instance, eliminating one or more category attributes) and at the level of statistical entity variable domains (for instance, selecting or compacting values). This process generally requires a calculation of the corresponding summary values; such a calculation is done according to rules which depend upon the summary type. The Mefisto approach to the management of the summary type is independent of the user, that is, the system "knows" if it is possible (and how) to compute the summary values with that particular summary type: the user applies the operators without having to consider the procedure by which the summary values are calculated. The Mefisto model does not consider the data analysis aspect: this activity requires the use of statistical packages and programming languages, according to the type of data analysis required.

Formally we have: let $\mathfrak{U} = \{C_1, \ldots, C_m\}$ be a *universe of category attributes* (composed of atomic and set-valued attributes).
Let *dom* be a function that associates to each category attribute C_i an underlying domain, $dom(C_i)$.

A *statistical entity scheme S* is a pair $< \underline{C}, t >$, where "$\underline{C}$" ($\mathfrak{U} \supseteq \underline{C}$) is a set of the category attributes of S and "t" is the summary type of the summary attribute, that is the abstract characterization of the g function (explained below).
Let $S = < \underline{C}, t >$ be an statistical entity scheme, with $\underline{C} = \{C_1, \ldots, C_n\}$ and $n \leq m$: a *statistical entity s* on S is a pair $< \underline{D}, g >$, where $\underline{D} = \{D_i\}$, $1 \leq i \leq n$, $dom(C_i) \supseteq D_i$ and D_i finite set, and g is a function (according to t) which maps from the *statistical entity space* $r^+(s) = D_1 \times D_2 \times \ldots \times D_n$ to $\Re \cup \{N.A.\}$, where $\Re$ is the set of real numbers and N.A. corresponds to "Not Available".
We denoted by r (*s*) a relation which is a subset of the statistical entity space $r^+(s)$, that is, $r^+(s) \supseteq r(s)$. The relations r (*s*) and $r^+(s)$ are defined on the relation scheme $\underline{C}$ (that is, the set of category attributes previously defined). We represent an element of $r^+(s)$ by a tuple $p = (d_1, \ldots, d_n)$, with $d_i \in D_i$ $(1 \leq i \leq n)$, and let Z be a non empty subset of $\underline{C}$, that is, $\underline{C} \supseteq Z$; we denote by $p(C_i)$ the instance of attribute C_i ($C_i \in \underline{C}$) in the tuple p, and by p[Z] the tuple, whose values are the corresponding instances, in p, of the attributes of Z.

In order to schematically represent the operations of the Mefisto model we use the graph in Fig.2, where the directed edge (representing an operation) goes from the input data structure to the output data structure.

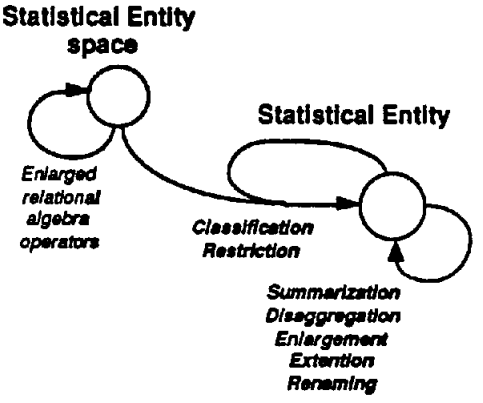

Figure 2

The algebra operations can have one or two SEs in input, or the pair < relation, SE >; output is one SE.
Relations in non-first normal-form (that is, relations having set as tuple components [OzOM 87]) are needed to manipulate the statistical entity space.

Now we illustrate briefly the most important operations of the Mefisto algebra.

Summarization

One of the operations performed in the statistical entity manipulation is the elimination of one category attribute. This operation is carried out by the summarization operation, which provides as output a statistical entity in which the category attributes are the same (except the deleted category attribute) as those of the input statistical entity; the summary values are computed according to the new statistical entity space of the output statistical entity.

For instance, let us consider the statistical entity "number_of_cars_produced" of Fig.3 described by "model" and "years". If we wish to have the total quantity of cars produced described only by "years", we apply the summarization operation with respect to the category attribute "model", obtaining the SE of Fig.4.

number of car produced in Japan *absolute value*		model		
		Corolla	Civic	Corona
years	1980	427	341	220
	1981	458	373	249
	1982	499	401	285

Note: the summary values are expressed in thousands.

Figure 3

total of car produced in Japan

absolute value

years	1980	988
	1981	1080
	1982	1185

Note: the summary values are expressed in thousands.

Figure 4

Let $s_1 = < D_1, g_1 >$ be a statistical entity defined on $S_1 = < C_1, t >$ and let $C_i \in C_1$. The **summarization** of s_1, with respect to C_i, is the statistical entity s defined on $< C_1 - \{C_i\}, t >$; then we have:

$$s = \Sigma_{C_i}(s_1) = < D_1 - \{D_i\}, g >$$

with each summary value:

$$g(p) = f_t (\{ g_1(p_1) \mid p_1 \in r^+(s_1) \wedge p = p_1[C_1 - \{C_i\}] \})$$

where $p \in r^+(s)$.

f_t represents a function which depends on the summary type of the SE and which permits the computation of the summary values. This means that the summarization operation is actually a *family* of operations, each of them doing the same operation at the metadata level (the category attribute deletion) but by different f_t.

In the above example if the summary type was "average", the function f_t is different, because, obviously, the sum of the averages generally is not the same type of average.

Classification

This operation classifies one category attribute of an SE according to a given relation in which the new classification criteria are specified. For example, the category attribute "months" can be classified in "quarter" specifying the function of transformation (in this case "{January, February, March}" -> "1st quarter", etc).
This operation is also able to reclassify by means of a substitution a set (or possibly all) of the category attributes of the SE with another set of attributes, according to a given relation. The resulting statistical entity is a reorganization of the starting data, with calculation of the summary values.
Let us consider, for example, the statistical entity "number_of_ cars_produced_in_Japan" of Fig. 3; if we wish to have the same statistical entity described by the category attributes "displacement" and "years" (the link between the "model" and "displacement" is the relation "rel" of Fig. 5-a), we perform the classification of the above SE using the relation "rel" along the category attributes "displacement" and "years" and we obtain the statistical entity of Fig. 5-b.

rel

model	displacement
Corolla	1,2
Civic	1,2
Corona	1,8

Figure 5-a

the car produced in Japan *absolute value*		displacement 1,2	1,8
years	1980	768	220
	1981	831	249
	1982	900	285

Note: the summary values are expressed in thousands.

Figure 5-b

Let $s_1 = < D_1, g_1 >$ be a statistical entity on $S_1 = < C_1, t >$; let R be a relation scheme; $C_1 \cap R$ functionally determines R;. Let C be a set of attributes such that $C_1 \cup R \supset C$; let r

be a relation defined on $\underline{R}$. The **classification** of s_1 by r along $\underline{C}$ is a statistical entity s defined on $S = < \underline{C}, t >$:

$$s = \zeta_{\underline{C}}(s_1, r) = <\{ D_i \mid D_i \in \underline{D}_1 \wedge C_i \in \underline{C}_1 \cap \underline{C} \} \cup \{ D_i \mid D_i = \pi_{C_i}(r) \wedge C_i \in \underline{R} - \underline{C}_1 \}, g>$$

with summary value:

$$g(p) = f_t (\{ g_1(p_1) \mid p_1 \in r^+(s_1) \wedge p_r \in r \wedge p_1 [\underline{C}_1 \cap \underline{C}] = p [\underline{C}_1 \cap \underline{C}] \wedge p_r [\underline{R} \cap \underline{C}] = p [\underline{R} \cap \underline{C}] \wedge p_1 [\underline{R} - \underline{C}] = p_r [\underline{R} - \underline{C}] \})$$

where $p \in r^+(s)$.

Also, if $\{g_1(p_1)\} = \emptyset$ then $g(p) = f_t(\emptyset)$ = null value.

Restriction

This operation produces an output SE whose statistical entity space is restricted to the elements of a set described by a given relation. The result of its application is a new statistical entity, defined on the same statistical entity scheme as the input SE.
For instance, let us consider the SE "Employees" described by "industry" and "state" (Fig. 6); if we only want to have the distribution in "Florida, Texas", we perform the restriction of "Employees" by the relation "rel" of Fig. 7, obtaining the SE of Fig. 8.
We note that there is the calculation of the summary values, because the summary type is "rate".

employees		industry			
rate per industry, state		agricolture	metal	other	*total*
state	California	1,2	5,4	9,1	*15,7*
	Florida	2,1	10,1	12,1	*24,3*
	Oregon	2,4	11,5	7,5	*21,4*
	Texas	6,4	12,7	19,5	*38,6*
	total	*12,1*	*39,7*	*48,2*	*100*

Figure 6

rel
state
Florida
Texas

Figure 7

employees in the South		industry			
rate per industry, state		agricolture	metal	other	*total*
state	Florida	2,89	16,22	19,55	*38,66*
	Texas	9,67	20,17	31,5	*61,34*
	total	*12,56*	*36,39*	*51,05*	*100*

Figure 8

Let $s_1 = < \underline{D}_1, g_1 >$ be an SE on $S_1 = < \underline{C}_1, t >$; further, let r_1 be a relation defined on the relation scheme $\underline{R}_1$ (this relation represents a set of element of the statistical entity space of s_1) and let $\underline{C}_1 \supseteq \underline{R}_1$ and $\pi_{\underline{R}_1}(r^+(s_1)) \supseteq r_1$. The **restriction** of s_1 by r_1 is a statistical entity s defined on the same statistical entity scheme S_1:

$$s = \sigma_M(r_1, s_1) = < \{ D_i \mid D_i = D_{i1} \wedge C_{i1} \in \underline{R}_1 \} \cup \{ D_i \mid D_i = \pi_{C_i}(r_1) \wedge C_i \in \underline{R}_1 \}, g >$$

with summary value:

$$g(p) = f_t (g_1(p_1) \mid p_1 \in r^+(s_1) \wedge p_1 = p)$$

where $p \in r^+(s)$.

Enlargement

Let us consider, as an example, the two SEs described by "industry" and "state": "Employees in the South" where the statistical variable domain of "state" is {Florida, Texas} (Fig. 8); "Employees in the West" where the the statistical variable domain of "state" is {Oregon, California} (Fig. 9). If we are interesting to have the SE "Employees" described by "industry" and "state" where the statistical variable domain of "states" is {Oregon, Florida, California, Texas}), that is, the "union" of the previous two SEs, we perform the enlargement of these SEs and we obtain the SE of Fig. 6 (where the weight of the two input SEs with respect to the output one is <40, 60>).

Let $s_1 = < \{D_1', D_2, \ldots, D_n\}, g_1 >$ and $s_2 = < \{D_1'', D_2, \ldots, D_n\}, g_2 >$ be two statistical entities defined on $S_1 = < \underline{C}_1, t >$, so that $D_1' \cap D_1'' = \emptyset$. The **enlargement** of s_1 and s_2 is a statistical entity s defined on the same statistical entity scheme S_1:

$$s = s_1 \Omega s_2 = < \{ D_1' \cup D_1'', D_2, \ldots, D_n \}, g >$$

with summary value:

$$g(p) = \begin{cases} f_t(g_1(p_1)) & \text{if } p_1 \in r^+(s_1) \wedge p_1 = p \\ f_t(g_2(p_2)) & \text{if } p_2 \in r^+(s_2) \wedge p_2 = p \end{cases}$$

where $p \in r^+(s)$.

employees in the West		industry			
rate per industry, state		agricolture	metal	other	*total*
state	California	3	13,5	30,2	*46,7*
	Oregon	5,85	28,7	18,75	*53,3*

Figure 9

4 Discussion

The case of macro data is more complex than that of micro data. Generally simple queries for manipulating SE are expressed in a complex way (especially when the summary type is not "absolute value") both by classic query languages, such as Sql, and by some proposals such as the Extended Relational Algebra proposed by [OzOM 87] (it is also an extension of Klug's algebra [Klug 82]). Even if the extension is different [Su 83] and the query is expressed in a simpler way, it is difficult to use the operations, because there are some conditions of applicability for each operations: for example in the case of projection there are three conditions of applicability; this fact causes difficulty in their use.
The fact that the models based on the relational model are an appropriate tool for the logical representation and manipulation of micro-DBs, but have these disadvantages for the case of macro data, should not be surprising. In fact, at the level of micro data, the only things that need to be represented are the concepts and the associations among such concepts: on the other hand, the generation of macro data from micro data renders such associations implicit to the summary datum; we must express, for each summary attribute, the category attributes that define it and viceversa [ChSh 81].
The Mefisto model has some advantages with respect to the models which currently exist in literature, in particular the relational model. It represents explicitly the link between the category attributes and a summary attribute, that is, the summary attribute is a function of category attributes; for example, the SE in Fig. 1 is denoted by:

cancer_around_the_world (sex, country, site)

Also, the queries are expressed in a simple way: for example, if the user wishes to know "what is the distribution of cancer by country and by site", he performs the following command expressed in Staquel [MeoE 90], the language based on the Mefisto model:

$$\Sigma_{sex}\ \text{cancer\_around\_the\_world}$$

The operations must at least have the power of matrix algebra in order to allow data analysis. In Mefisto, it is possible define with the same approach the operations needed for data analysis
As an example we will define the division operation which performs the classic arithmetic operation of *divide*: let $s_1 = < \{D_1', D_2', \ldots, D_h', \ldots, D_n'\}, g_1 >$ be a statistical entity defined on $S_1 = < \underline{C}_1, t_1 >$ and $s_2 = < \{D_1'', D_2'', \ldots, D_h''\}, g_2 >$ be a statistical entity defined on $S_2 = < \underline{C}_2, t_2 >$, such that $\underline{C}_1 \supseteq \underline{C}_2$ and for $1 \le i \le h$, $D_i'' = D_i'$.

The division of s_1 by s_2 is a statistical entity s defined on the scheme $S = < \underline{C}_1, t >$:

$$s = s_1 + s_2 = < \{D_1', D_2', \ldots, D_n'\}, g >$$

with summary value:

$$g(p) = (g_1(p_1) / g_2(p_2) \mid p_1 \in r^+(s_1) \wedge p_2 \in r^+(s_2) \wedge p_1 = p \wedge p_2 = p[\underline{C}_2])$$

where the symbol / represents the division between real numbers and $p \in r^+(s)$.

Let us take as an example the SE in Fig. 3; if we wish to obtain the SE which expresses the "percentages with respect to the total per year", we must apply the division operation between the SEs of Fig. 3 and of Fig. 4 (obtained as a summarization with respect to "model" of SE Fig. 3). The division gives as a result the SE in Fig. 10.

percent of car produced in Japan		model		
rate per year		Corolla	Civic	Corona
years	1980	0,44	0,34	0,22
	1981	0,42	0,35	0,23
	1982	0,42	0,34	0,24

Figure 10

It must be realized that it is impossible a priori to know which summary types should be considered in a macro-DB, since there is always the possibility of defining new ones. The database administrator must supply the system, for each summary type, the information (expressed in general terms and not tied to the single SE) which enables the system to construct the computation procedures of the summary values for the summarization, the classification, the restriction and the enlargement operations. This is a problem of knowledge elicitation. Consequently the object oriented model is used as the knowledge structure for capturing the statistical expertise, needed to compute the summary values; in [Falc 89] a model based on objects was defined: it uses a double inheritance hierarchy (between classes and between instances).

5 Conclusions

In this paper we have presented the Mefisto model, based on the functional approach, in which the *statistical entity* structure and a set of *new operators* for SE manipulation are proposed and discussed.
Prototype implementations of a visual Data Definition Language, based on the graphical model (Grass [RaRi 83]), and of a Query Language [MeoE 90], as well as of Visual end-user interface [RaRi 90] have been implemented on a Macintosh II.

Future work deals with the following problems:
- design and implementation of an integrated environment for the DDL, the QL and the data analysis activity;
- improvement of the current proposal regarding the use of an object-oriented approach [Falc 89], in order to extend the above data analysis to complex statistical indicators.

Acknowledgement

The authors wish to thank A. Segev for precious suggestions and useful discussions, and F. Ferri, P. Grifoni and L. Meo Evoli for the implementation.

References

[BrNS 83] Brown V.A., Navathe S.B., Su S.Y.W. "Complex data types and data manipulation language for statistical and scientific databases", Proceed. of the II° Intern. Workshop on Statistical Database Management, Los Altos, Ca, Sept. 1983

[ChSh 81] Chan P., Shoshani A. "SUBJECT: a directory driven system for organizing and accessing large statistical databases", Proceed. 7th Very Large Data Bases, Cannes, France, Sept. 1981

[Falc 89] Falcitelli G., Meo Evoli L., Nardelli E., Ricci F.L. "The Mefisto* model: an object oriented representation for statistical data management" ,in "Data Analysis, Learning Symbolic and Numeric Knowledge", Nova Science Publishers, 1989

[Ghosh 86] Ghosh S.P. "Statistical relational tables for statistical database management", IEEE Transactions on Software Engineering, vol.SE-12, no.12, 1986

[Klug 82] Klug A. "Equivalence of relational algebra and relational calculus query languages having aggregate functions", Journal ACM, Vol.29, No.3, July 1982

[MeoE 90] Meo-Evoli L. "Interaction model for Man-Table Base System", Proceed. of the VII IASTED International Symposium on Applied Informatics, Innsbruck, Austria, Feb., 1990.

[OzOM 87] Ozsoyoglu G., Ozsoyoglu Z.M. , Matos V. "Extending Relational Algebra and Relational Calculus with Set-Valued Attributes and Aggregate Functions", ACM Transaction on Database Systems, Vol.12, No.4, Dec. 1987

[RaRi 83] Rafanelli M., Ricci F.L. "Proposal of a logical model for statistical database", Proceed. of the II° Intern. Workshop on Statistical Database Management, Los Altos, Ca, Sept. 1983

[RaRi 90] Rafanelli M., Ricci F.L. "A visual interface for brawsing and manipulating statistical entities", Lecture Notes in Computer Science, 420, Springer Verlag Pub., 1990

[Shos 82] Shoshani A. "Statistical databases: characteristics, problems and solutions", Proceed. 8th Very Large Data Bases, Mexico City, Mexico, Sept. 1982

[ShWo 85] Shoshani A., Wong H.K.T. "Statistical and scientific database issues", IEEE Transactions on Software Engineering, Vol.SE-11, No.10, Oct. 1985

[Su 83] Su S.Y.W. "SAM*: a semantic association model for corporate and scientific / statistical databases" Information Sciences, Vol.29, No.2-3, May - June 1983

[Wong 84] Wong H.K.T. "Micro and macro statistical / scientific database management", Proceed. 1st IEEE Intern. Conference on Data Engineering, Los Angeles, Ca, Apr. 1984