Extending the E/R Model for the Multidimensional Paradigm

Carsten Sapia, Markus Blaschka, Gabriele Höfling, Barbara Dinter

FORWISS (Bavarian Research Center for Knowledge-Based Systems)
Orleansstr. 34, D-81667 Munich, Germany
Email: {sapia, blaschka, hoefling, dinter}@forwiss.tu-muenchen.de

Abstract. Multidimensional data modeling plays a key role in the design of a data warehouse. We argue that the Entity Relationship Model is not suited for multidimensional conceptual modeling because the semantics of the main characteristics of the paradigm cannot be adequately represented. Consequently, we present a specialization of the E/R model - called Multidimensional Entity Relationship (ME/R) Model – that is suitable for the conceptual modeling of OLAP applications. In order to express the multidimensional structure of the data we define two specialized relationship sets and a specialized entity set. The resulting ME/R model allows the adequate conceptual representation of the multidimensional data view inherent to OLAP, namely the separation of qualifying and quantifying data and the complex structure of dimensions. We demonstrate the usability of the ME/R model by an example taken from an actual project dealing with the analysis of vehicle repairs.

1 Introduction

Multidimensional data modeling plays a key role during the design of a data warehouse. The multidimensional warehouse schema offers an integrated view on the operational data sources. Consequently, it serves as the core of the data warehouse and as the basis for the whole warehouse development and maintenance cycle. Due to this central role sufficient attention should be paid to the development of this schema. Figure 1 sketches the process of the schema design in data warehousing environments. The schema is mainly influenced by user requirements and the availability and structure of the data in operational systems. Most warehousing projects take an evolutionary approach[1], i.e. start with a prototype providing a certain functionality and set of data. This prototype will be further adopted according to the changing and growing requirements gained from users' feedback. Thus, in warehouse maintenance, the user requirements are subject to frequent changes making schema evolution an important issue. To assure the flexibility and re-usability of the schema in such an environment, the model must be specified on a conceptual level (e.g. using the Entity Relationship

[1] both in our experience from industrial projects [9] and in the warehouse literature, see e.g. [10]

Model). This means especially that it must not assume any facts that are the result of further design steps e.g. the decision which database technology is to be used (multi-dimensional vs. relational).

For OLAP and data warehouse systems this is even more important as the most common design methodologies mix up the conceptual and the logical/physical design. Currently the state of art in dimensional modeling is the use of implementation (mostly even tool specific) formalisms for data modeling. For example, the ubiquitous star schema is not conceptual in the sense that it assumes the relational implementation and contains further decisions (e.g. denormalization) that should be subject of the physical design phase.

There is a consensus ([10], [12], [14]) that the multidimensional paradigm comes very close to the inherent structure of the problem domain (decision support systems). In this paper we investigate the special requirements of the multidimensional paradigm. We argue that the established conceptual design methods used for relational (e.g. the Entity Relationship Model [4]) or object-oriented systems do not offer the necessary support to reflect the multidimensional data model in a natural and intuitive way. Moreover, some of the multidimensional semantics is lost when expressing a multidimensional schema with these techniques. This means that the semantics must be represented informally which makes them unusable for the purpose of automatic generation (e.g. automatic generation of database schemes or query tools).

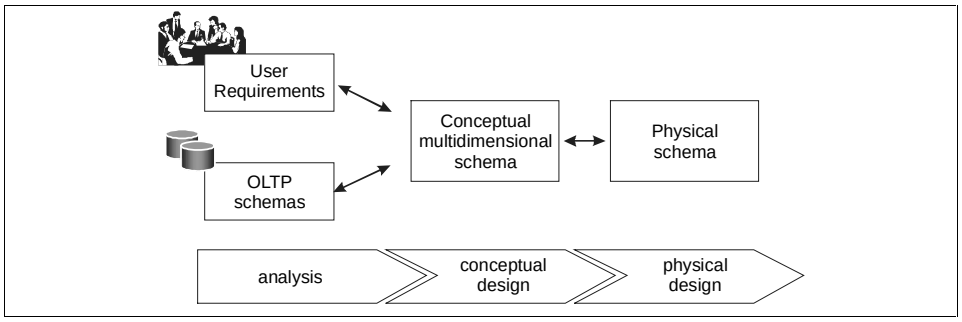

Fig. 1. Schema design process in data warehousing environments

Thus, a conceptual multidimensional model capable of expressing the multidimensional semantics is necessary. However, the scientific community and the vendors are still debating about the formal definition of the multidimensional model and its properties. Each product respectively author presents a model of different expressiveness.

Possible approaches to an expressive conceptual multidimensional model are to build a new model from scratch (which also means additional effort for its formal foundation) or to use an existing, general-purpose model and modify it so that the special characteristics of the multidimensional paradigm can be expressed.

Consequently, this paper presents a multidimensional specialization of the E/R model - called Multidimensional E/R Model (ME/R Model). By basing our approach on an established model we enable the transfer of the research results published in the context of the E/R model. This includes especially the work about automatic schema

generation and formal foundation of the semantics. Furthermore, it is possible to make use of the proven flexibility of the well accepted E/R model.

The remainder of this paper is structured as follows: section 2 informally introduces the multidimensional paradigm and states the special requirements of OLAP applications regarding the data model. Section 3 describes the specializations of the E/R model that are necessary to fulfil these requirements and defines the ME/R model. In section 4 we investigate the expressive power of the ME/R model. To demonstrate the feasibility of our approach we model a real world example (section 5). Finally, we present related (section 6) and future work (section 7).

2 The Multidimensional Paradigm

The multidimensional paradigm is useful for a multitude of application areas (e.g. GIS, PACS, statistical databases and decision support). For the purpose of this paper we focus on typical OLAP applications. For example a vehicle manufacturer might want to analyze the vehicle repairs to improve his product, define new warranty policies and to get information about the quality of the garages.

Often a cube metaphor ([3]) is used to represent this data view as shown in figure 2. Such a cube corresponds to a subject of analysis called *fact* (e.g. vehicle repair). The cells of the data cube contain the (mostly numerical) *measures* (also called *quantifying data*) describing the fact (e.g. costs and duration of the vehicle repair). The axes of the cube (called *dimensions* or *qualifying data*) represent different ways of analyzing the data (e.g. vehicle and time of the repair).

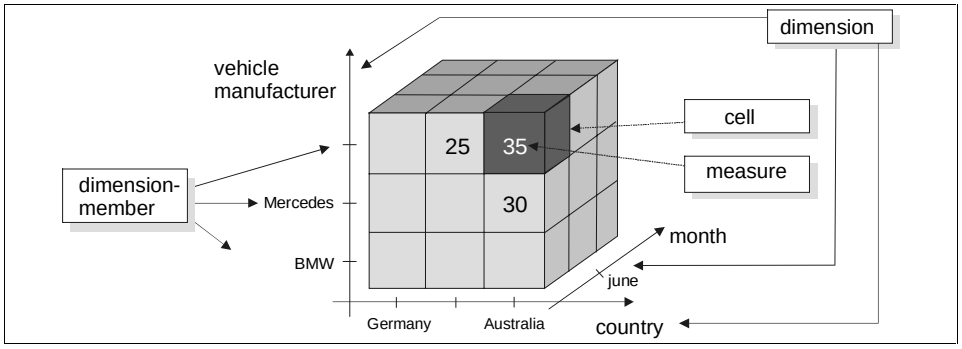

Fig. 2. A visualization of a multidimensional schema using the cube metaphor

This data view is similar to the notion of arrays. However, with arrays the dimensions of the multidimensional data space are only structured by a linear order defined on the indexes. For OLAP applications this is not sufficient because from the point of view of the OLAP end-user, the elements (respectively instances) of an OLAP dimension (called dimension members) are normally not linearly ordered (e.g. garages)[2].

[2] A prominent exception to this rule is the time dimension that possesses an inherent order

Instead classification hierarchies containing levels are used for the structuring of dimensions. A hierarchy level contains a distinct set of members. Different levels correspond to different data granularities (e.g. daily figures vs. monthly figures) and ways of classification (e.g. geographic classification of garages vs. classification of garages by type). Level A rolls up to a level B if a classification of the elements of A according to the elements of B is semantically meaningful to the application (e.g. the level 'days' rolls up to 'month').

A level can roll up to any number of levels thus forming multiple hierarchies on a single dimension. This case occurs if different criteria of classification are possible for dimension members. For example, garages can be classified by their geographical location and their type (see example in section 5). Another special case of hierarchies are alternative paths. This type of hierarchy occurs if several rolls-up paths exist between two levels. An example for this is the classification of cities by geographical regions and federal districts (see figure 3).

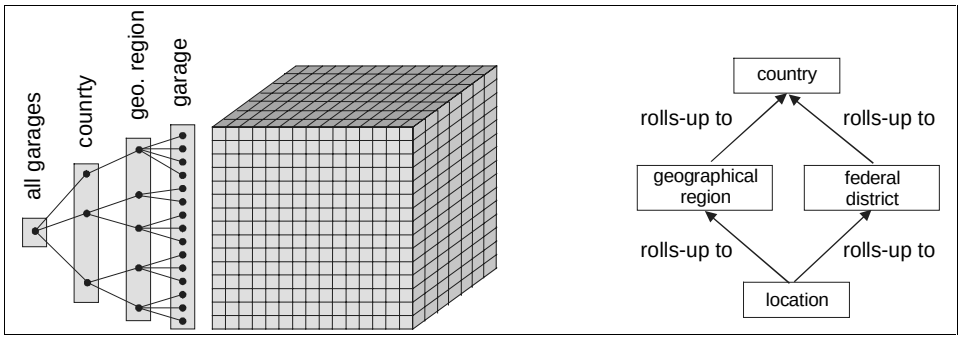

Fig. 3. Hierarchy levels structure the dimensions(left). Alternative pathes within a dimension(right)

Another orthogonal way of structuring dimensions from a users point of view is the use of dimension level attributes. These attributes describe dimension level members but do not define hierarchies (e.g. the name and address of a customer or the name of the region manager).

Not only qualifying data but also quantifying data possesses an inherent structure. In most applications different measures describing a fact are common (e.g. for a vehicle repair it might be useful to measure the duration of the repair, the costs for parts being exchanged and the cost for the wages). That means that a cell of the cube does contain more than one numeric value. Some of these measures are derived, i.e. they can be computed from other measures and dimension attributes (e.g. total repair cost is the sum of part costs and costs for wages).

A complex schema can contain more than one cube. This becomes necessary if an application requires the analysis of different facts (e.g. vehicle sales and repairs) or if not all of the measures are dependent on the same set of dimensions.

These multiple cubes can share dimensions (e.g. the time dimension). This does not necessarily mean that these cubes measure data using the same granularity. For example vehicle sales might be recorded and analyzed on a weekly basis, while the vehicle repairs are recorded daily (see section 5).

Regarding the multidimensional paradigm it is obvious, that the E/R model is not very well suited for the natural representation of multidimensional schemas. The inherent separation of qualifying and quantifying data cannot be expressed as all entity sets are treated equally by the E/R model. Furthermore the semantics of the complex structure of the dimensions (rolls-up relationship between dimension levels) is an integral part of the multidimensional paradigm that is too specific to be modeled as a general purpose relationship.

3 The Multidimensional E/R Model

In order to allow the natural representation of the multidimensional semantics inherent to OLAP schemas, the E/R model is specialized. Of course, there are several possible ways to achieve this goal. Our design was driven by the following key considerations:

- *Specialization of the E/R model*: All elements that are introduced should be special cases of native E/R constructs. Thus, the flexibility and expressiveness of the E/R model is not reduced.
- *Minimal extension of the E/R model*: The specialized model should be easy to learn and use for an experienced E/R modeler. Thus, the number of additional elements needed should be as small as possible. A minimal set of extensions ensures the easy transferability of scientific results (e.g. formal foundations) from the E/R model to the ME/R model by discussing only the specific extensions.
- *Representation of the multidimensional semantics*: Despite the minimality, the specialization should be powerful enough to express the basic multidimensional semantics, namely the separation of qualifying and quantifying data and the hierarchical structure of the qualifying data.

A lot of variations of the E/R model (for an overview see e.g. [17]) have been published since the first proposal of Chen. For the purpose of this paper we use a very basic version of the E/R model. We formally describe our specialized E/R model using the meta modeling approach. We adhere to the four layer technique of the ISO/IRDS standard for metadata [11]. Figure 4 shows the meta model of our M/ER model (Dictionary Definition Layer of the IRDS). The part with the white background shows the meta model of the E/R model we use as a foundation. For the purpose of describing the meta model, we make use of an extended version of the E/R model which allows the concept of generalization. This is done to increase the readability of the meta model. However, the decision which type of constructs are allowed in the E/R model itself (and thus the ME/R model) is left open to the modeler.
Following our key considerations we introduce the following specialization:

- a special entity set: dimension level,
- two special relationship sets connecting dimension levels:
  - a special n-ary relationship set: the '*fact*' relationship set and
  - a special binary relationship set: the '*rolls-up to*' relationship set.

Since the semantic concept ‘dimension level’ is of central importance, we introduce a special entity set for dimension levels.

To model the structure of qualifying data we introduce a special binary relationship set: the rolls-up relationship. It relates a dimension level A to a dimension level B representing concepts of a higher level of abstraction (e.g. city *rolls-up* to country). The rolls-up graph is defined as follows: *RG* = (E,V) with E being the finite set of all dimension levels $e_1,..,e_k$ and V = { $(e_i,e_j) \mid i \neq j \wedge 1 \leq i,j \leq k \wedge e_i$ *rolls-up to* e_j }. Due to the special semantics of the roll up relation, no cycles must be contained in the graph as this could lead to semantically not reasonable infinte roll-up pathes (e.g. day rolls-up to month and month rolls-up to day). This means the following global integrity constraint must be fulfilled ($\rightarrow^*$ denotes the transitive closure of the *rolls-up* relation):

$$\forall\ e_i, e_j \in E : e_i \rightarrow^* e_j \Rightarrow i \neq j$$

Thus the rolls-up graph *RG* is a directed acyclic graph (DAG). The name attribute of the roll-up relation set describes the criteria of classification. (e.g. ‘lives in’ for the roll-up relationship set connecting ‘customer’ and ‘geographical region’)

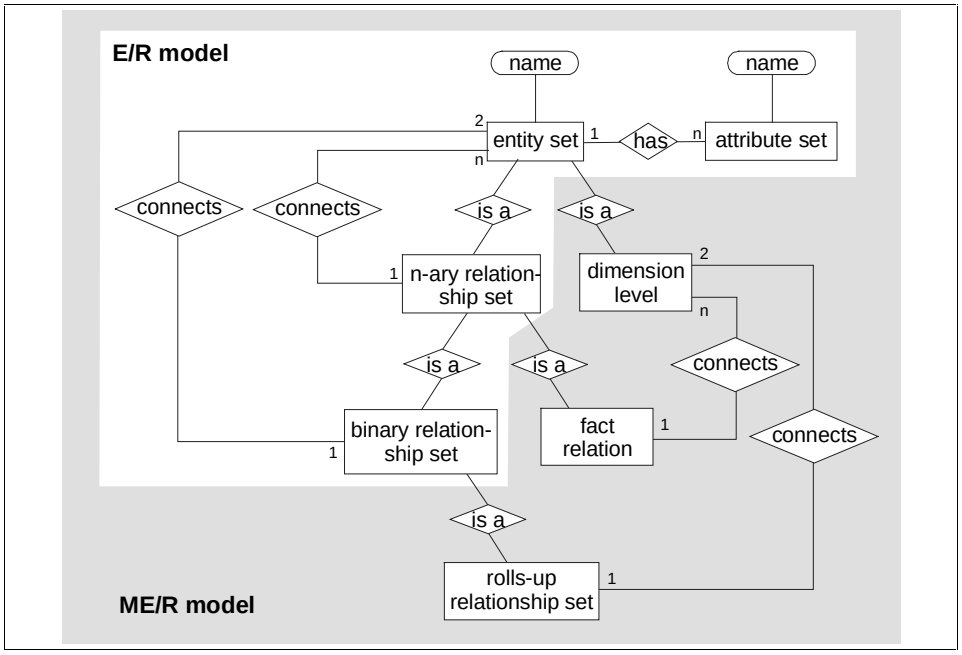

Fig. 4. The meta model of the ME/R model is an extension of the meta model of E/R.

The fact relationship set is a specialization of a general n-ary relationship set. It connects n different dimension level entities. Such a relation represents a fact (e.g. vehicle repair) of dimensionality n. A description of the fact is used as the name for the set. The directly connected dimension levels are called *atomic* dimension levels.

The fact relationship set models the inherent separation of qualifying and quantifying data. The attributes of the fact relationship set model the measures of the fact (quantifying data) while dimension levels model the qualifying data.

To distinguish our specialized elements from the native E/R modeling elements and to enhance the understandability of the graphical model, we use a special graphical notation for dimension level sets, fact relationship sets, and rolls up relationship sets (figure 5).

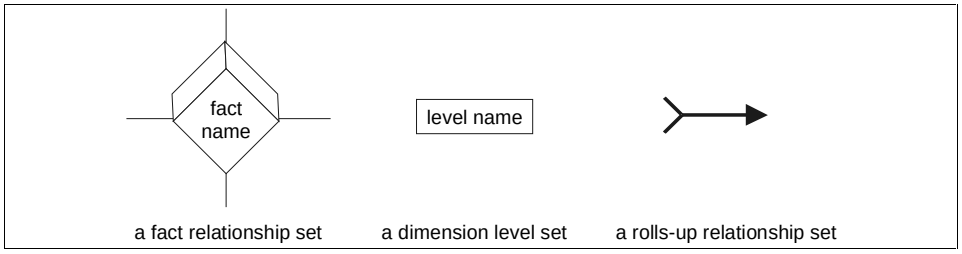

Fig. 5. The graphical notation of the ME/R elements.

4 Distinctive Features of the ME/R model

After having introduced the ME/R model, we now investigate how the ME/R model fulfills the requirements of the multidimensional paradigm. An example for modeling a real-world scenario can be found in the next section.

A central element in the multidimensional paradigm is the notion of dimensions that span the multidimensional space. The ME/R model does not contain an explicit counterpart for this concept. This is not necessary because a dimension consists of a set of dimension levels. The information which dimension-levels belong to a given dimension is included implicitly within the structure of the rolls-up graph. Formally, the fact relationship identifies the n atomic dimension levels $e_{i1},..e_{in}$. The according dimensions D_k are the set of the dimension levels that are included in the subgraph of the rolls-up graph *RG(E,V)* defined by the atomic level.

$$D_k = \left\{ e \in E \mid e_{i_k} \rightarrow^* e \right\} \quad 1 \le k \le n$$

The hierarchical classification structure of the dimensions is expressed by dimension level entity sets and the roll-up relationships. As previously noted, the rolls-up relationship sets define a directed acyclic graph on the dimension levels. This enables the easy modeling of multiple hierarchies, alternative paths and shared hierarchy levels for different dimensions (e.g. customer and garage in figure 7). Thus no redundant modeling of the shared levels is necessary. Dimension level attributes are modeled as attributes of dimension level entity sets. This allows a different attribute structure for each dimension level.

By modeling the multidimensional cube as a relationship set it is possible to include an arbitrary number of facts in the schema thus representing a ‘multi-cube model’. These different cubes and their shared dimensions can be expressed as shown

in figure 7. Notably the schema also contains information about the granularity level on which the dimensions are shared. This information is for example necessary for the design of multidimensional joins during further development steps.

Regarding measures and their structure the ME/R model allows record structured measures as multiple attributes are possible for one fact relationship set. The semantic information that some of the measures are derived cannot be included in the model. Like the E/R model the ME/R model captures the static structure of the application domain. The calculation of measures is a functional information and should not be included in the static model. An orthogonal functional model should capture these dependencies.

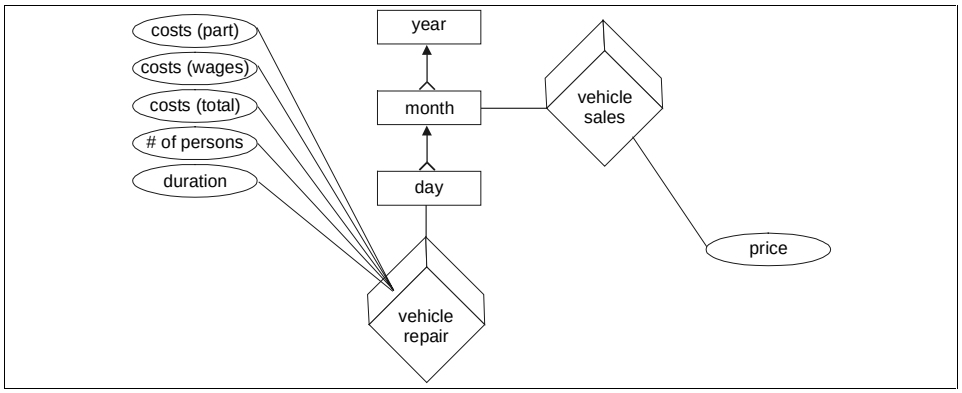

Fig. 6. Multiple cubes sharing a dimension on different levels

5 Applying the ME/R Model (Case Study)

To demonstrate the feasability of our ME/R model, we present a real application. The following example is taken from a project with an industrial partner [9]. An automobile manufacturer stores data about repairs of vehicles. Among other, the date of repair, properties of the vehicle (e.g. model), information about the specific repair case (e.g. costs, number of garage employees involved, duration of the repair), data about the garage doing the repair, and data about the customer who owns the vehicle are stored.

Typical examples queries for this scenario are:

- "Give me the average total repair costs per month for garages in Bavaria by type of garage during the year 1997"
- "Give me the five vehicle types that had the lowest average part costs in the year 1997"

The first design step is to determine which data are dimensions and which are facts. We assume that the repair costs (broken down by part costs, wages and total) for a specific vehicle (owned by a customer) for a specific garage are given on a daily basis. Then the facts (quantifying data) are the repair costs (parts, wages, total). Vehi-

cle, customer, garage and day are the corresponding dimensions (qualifying data) and because they are at the finest granularity also the atomic dimension levels. Thus, the *fact* relationship connects the *vehicle repair fact* with the dimensions vehicle, customer, garage and day. The *rolls-up* relationships are shown in figure 7 which contains the complete ME/R diagram for this case study. The *fact* relationship in the middle of the ME/R diagram connects the atomic dimension levels. Each dimension is represented by a subgraph that starts at the corresponding atomic level (e.g. the time dimension starts at the dimension level day and comprehends also month and year). The actual facts (part costs, wages, total costs, number of persons involved and duration of the repair) are modeled as attributes of the *fact* relationship. The dimension hierarchies are depicted by the *rolls-up* relationships (e.g. vehicle *rolls-up* to model and brand). Additional attributes of a dimension level (e.g. age or income of a customer) are depicted as dimension attributes of the corresponding dimension level.

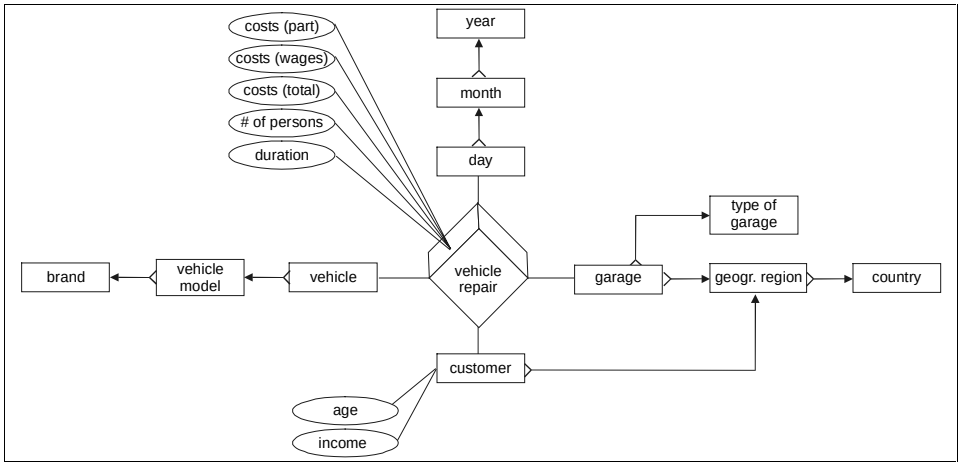

Fig. 7. The ME/R diagram for the analysis of vehicle repairs

Notably, the schema contains a *rolls-up* relationship between the entities *'customer'* and *'geographic region'* and between *'garage'* and *'geographic region'*. This shows a distinctive feature of our model: levels of different dimensions may roll up to a common parent level. This might imply that the dimensionality of the cube is reduced by one when executing the roll up. However, this is not the case as the model only captures the semantical fact, that the same type of classification (geographical classification) is used in both dimensions. During later phases of the development cycle this information can be used to avoid redundancies as the geographical regions only have to be stored once. The corresponding data cube however still contains two dimensions that contain the same members (customer geographical region and garage geographical region).

Since our ME/R model is a **specialization** of the E/R model, regular E/R constructs can also be used in ME/R diagrams. In our example, the entity *vehicle* can be extended e.g. to distinguish between cars and trucks. This scenario is shown in figure 8. We use the *isa* relationship to model the categorization of vehicles. The extended diagram (i.e. with 'regular' E/R constructs and special ME/R constructs) further al-

lows us to model additional features of the subtypes of our entity *vehicle*. Features [13] are attributes that are only meaningful in a subclass but not in a superclass. Different subclasses may have different features. For example, for a *vehicle* in general one might store attributes like *length*, *width, height, colour,* or *horse power*, but a feature like *loading capacity* or *loading area* in m² is only meaningful for trucks. For a car on the other hand, it might be useful to store the *number of seats* or the *type of gear* (i.e. manual or automatic). Thus, using these combined E/R and ME/R modeling technique, features as introduced in [13] can be modeled on a conceptual level.

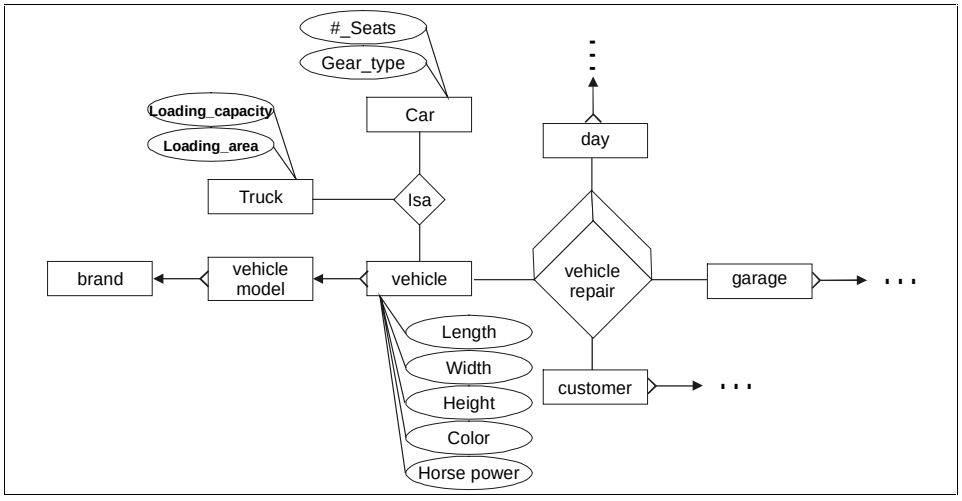

Fig. 8. Combining ME/R notation with classical E/R constructs

6 Related Work

A lot of publications are available concerning 'multidimensional data modeling'. Unfortunately only very few recognize the importance of the separation of conceptual and logical/physical issues. This is largely because the development in this area has so far been driven by the product vendors of OLAP systems. To our knowledge only very few papers investigating a graphical conceptual (i.e. implementation independent) data modeling methodology for multidimensional database have been published. Ralph Kimball proposes the design of Data Warehouses using a multidimensional view of the enterprise data. He presented a 'multidimensional modeling manifesto' [12]. However, his approach is not conceptual in the sense that it is not independent of the implementation (a relational implementation in the form of a 'star schema' is assumed).

In the area of statistical databases, graphical conceptual models to capture the structure and semantics of statistical tables have been proposed (e.g. [15], [16]) for a long time. The data warehouse research community focused mainly on physical issues (e.g. [8]) of data warehouse design. Quite a lot of work has also been done to formal-

ize the multidimensional data model (see [2] for a comparison) and to define query languages for these data models([1]). However, the formalisms are not suited for conceptual modeling of user requirements. Our work supplements these papers by providing a graphical conceptual layer (as the E/R model provides for the relational paradigm).

Nevertheless, recently the deficit in conceptual models has been recognized. [6] proposes a formal logical model for OLAP systems and showed how it can be used in the design process. The paper suggests a bottom-up approach to data warehouse design. The authors assume an integrated E/R schema of the operational data sources and give a methodology to transform this schema into a dimensional graph which can be translated into the formal MD model. Our model is more suited to a top-down approach modeling the user requirements independently from the structure of the operational systems.

[7] also proposes a conceptual model called dimensional fact (DF) scheme. In their paper they give a graphical notation and a methodology to derive a DF model form the E/R models of the data sources. Although the technique supports semantically rich concepts it is not based on a formal data model. Our approach is to specialize a well researched and formally founded model. Furthermore, the notation does not allow the modeling of alternative paths which we believe is an important requirement.

In [13] Lehner et al. present a conceptual multidimensional data model. They argue that the common classification hierarchies are not sufficient for all types of applications. Therefore, they propose feature descriptions as a complementary mechanism for structuring qualifying information. As the main focus of the paper is the extension of the paradigm, no graphical notation (apart from the cube visualization) is provided.

7 Conclusions and Future Work

We started from the fact that the multidimensional paradigm plays a central role in the data warehouse and OLAP design process. However the fundamental semantics of this paradigm cannot be adequately expressed using the E/R model. Consequently, we proposed ME/R, a specialization of the E/R model especially suited for the modeling of OLAP applications. We also defined a graphical notation for the new elements which allows intuitive graphical diagrams. Our technique allows the easy modeling of multidimensional semantics (namely the separation of qualifying and quantifying data and the complex structure of dimensions). Multiple hierarchies, alternative paths and shared dimension levels can be naturally expressed. By designing ME/R as a specialization of the common E/R model we ensure a shallow learning curve and a high intuitivity of the diagrams. Since the modeler can combine ME/R elements with classical E/R elements semantically rich models can be built. Finally, we demonstrated the flexibility and usefulness of our approach by modeling a real world example.

The ME/R model can serve as the core of a full scale data warehouse design methodology. Using the ME/R model it is possible to capture the multidimensional application semantics during the conceptual design phase of a data warehouse. This information can be used during later phases (physical design and implementation) of the data warehouse process. As the semantics are an integral formal part of the model

automatic and heuristic generation steps are possible (e.g. the generation of database schemes and optimization strategies). A first step in this direction would be the mapping of the ME/R model to the formal logical multidimensional data models that were proposed recently.

The ME/R model allows to capture the static data structure. The modeling of dynamical (e.g. anticipated query behavior) and functional (e.g. the additivity of measures along dimensions or the functional relationship between hierarchy levels) aspects deserve a deeper study. Currently we are investigating a dynamic and a functional model supplementing the static model (analogous to the OMT) and study the interrelationship between those models. Additionally, we are working on a classification of multidimensional schema evolution operations (e.g. 'add dimension level') and examine the impacts of these operations. To this end, we evaluate schema evolution approaches from object-oriented databases and investigate their feasibility in the multidimensional case.

References

[1] A, Bauer, W. Lehner: *The Cube-Query-Language (CQL) for Multidimensional Statistical and Scientific Database Systems,* Proc. of the 5th Conference on Database Systems for Advanced Applications, Melbourne 1997.

[2] M. Blaschka, C. Sapia, G. Höfling, B. Dinter: Finding Your Way through Multidimensional Data Models, DWDOT Workshop (DEXA 98), Vienna

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

[4] P.P-S. Chen: The Entity Relationship Model – Towards a Unified View of Data. ACM TODS Vol. 1, No. 1, 1976

[5] E. F. Codd: Extending the Database Relational Model to Capture More meaning. ACM TODS Vol. 4, No. 4 (December 1979)

[6] L. Cabibbo, R. Torlone: *A Logical Approach to Multidimensional Databases.* EDBT 1998.

[7] M. Golfarelli, D. Maio, S. Rizzi, *Conceptual design of data warehouses from E/R schemes*, Proc. 31st Hawaii Intl. Conf. on System Sciences, 1998.

[8] V. Harinarayan, A. Rajaraman, J. D. Ullman: *Implementing Data Cubes Efficiently.* Proc. SIGMOD Conference, Montreal, Canada, 1996

[9] G. Höfling, M. Blaschka, B. Dinter, P. Spiegel, T. Ringel: *Data Warehouse Technology for the Management of Diagnosis Data* (in German), in Dittrich, Geppert (eds.): *Datenbanksysteme in Büro, Technik und Wissenschaft* (BTW), Springer, 1997.

[10] W. H. Inmon: *Building the Data Warehouse*, 2nd edition, John Wiley &Sons, 1996

[11] IRDS Framework ISO/IEC IS 10027, 1990

[12] R. Kimball: *A Dimensional Modeling Manifesto,* DBMS Magazine, August 1997,

[13] W. Lehner, T. Ruf, M. Teschke: CROSS-DB: *A Feature-Extended Multidimensional Data Model for Statistical and Scientific Databases*, Proc. of the CIKM'96, Maryland.

[14] Micro Strategy Inc.: The Case For Relational OLAP, White Paper. 1995,

[15] M. Rafanelli, A. Shoshani: *STORM : A Statistical Object Representation,* SSDBM 90

[16] S.Y.W. Su : SAM*: *A Semantic Association Model for Corporate and Scientific-Statistical Databases,* in: Journal of Information Sciences 29, 1983

[17] T.J. Teorey: *Database Modeling and Design*, 2nd edition, Morgan Kaufmann 1994