EXTENDING THE OLAP FRAMEWORK FOR AUTOMATED EXPLANATORY TASKS

Emiel Caron[1], Hennie Daniels [1,2]
[1]Erasmus University Rotterdam, ERIM Institute of Advanced Management Studies, PO Box 90153, 3000 DR Rotterdam, The Netherlands, phone +31 010 4082574, e-mail: ecaron@fbk.eur.nl; [2]Tilburg University, CentER for Economic Research, Tilburg, The Netherlands

Abstract

The purpose of OLAP (On-Line Analytical Processing) systems is to provide a framework for the analysis of multidimensional data. Many tasks related to analysing multidimensional data and making business decisions are still carried out manually by analysts (e.g. financial analysts, accountants, or business managers). An important and common task in multidimensional analysis is business diagnosis. Diagnosis is defined as finding the "best" explanation of observed symptoms. Today's OLAP systems offer little support for automated business diagnosis. This functionality can be provided by extending the conventional OLAP system with an explanation formalism, which mimics the work of business decision makers in diagnostic processes. The central goal of this paper is the identification of specific knowledge structures and reasoning methods required to construct computerized explanations from multidimensional data and business models. We propose an algorithm that generates explanations for symptoms in multidimensional business data. The algorithm was tested on a fictitious case study involving the comparison of financial results of a firm's business units.

1. Introduction

In this paper, we describe an extension of the OLAP (On-Line Analytical Processing) framework with automated causal diagnosis, offering the possibility to automatically generate explanation and diagnostics to support business decision tasks. Today's OLAP systems have no explanation or diagnosis capabilities. Such functionality can be provided by extending the conventional OLAP system with an explanation formalism, which mimics the work of human decision makers in diagnostic processes. The formalisation of diagnostic problem-solving is a sub-area of Operations Research (OR) and Artificial Intelligence (AI). In [3] diagnosis is defined as finding the best explanation of observed abnormal behaviour of a system under study. Here we combine diagnostic problem solving and OLAP.

OLAP is defined as "*a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user*" [9]. The core component of an OLAP system is the data warehouse, which is a decision-support database that is periodically updated by extracting, transforming, and loading data from several OLTP (On-Line Transaction Processing) databases. An OLAP system or multidimensional model organizes data using the dimensional modelling approach, which classifies data into *measures* and *dimensions*. Measures or facts like, for example, sales figures and costs, are the basic units of interest for analysis. Measures represent countable or summable information concerning a business process. Dimensions correspond to different perspectives for viewing measures. Dimensions are usually organised as dimension hierarchies, which offers the possibility to view measures at different dimension levels (e.g. $month \prec quarter \prec year$). The hierarchies in a dimension specify the aggregation levels.

The objective of this paper is to extend the multidimensional model with an explanation formalism. For this purpose, the general model and methodology for automated

business diagnosis, as developed by Daniels and Feelders [3, 4], is applied to the multi-dimensional model. The definitions in the explanation model are adapted in order to connect them with measures and dimensions. In addition, an algorithm is described for maximal explanation of multidimensional data and applied on a case study.

The remainder of this paper is organized as follows. We first demonstrate the use of the explanation formalism in OLAP using an artificial dataset. Section 2 provides a short introduction to the OLAP framework and introduces the most important concepts of the multidimensional model, followed by an introduction to the causal explanation model in section 3. In section 4 the multidimensional model is extended with the explanation model in order to generate explanations for symptoms derived from multidimensional data. This section is followed by the introduction of an algorithm for maximal explanation in multidimensional data in section 5. We discuss some related work in section 6 and in section 7 the complete method is illustrated in a fictitious case study on business unit performance. Finally, conclusions are discussed in section 8.

1.1. Illustration

Here we will consider the dataset of the fictitious ABC-company and present an illustration of our OLAP explanation framework. The multidimensional dataset is composed out of the following financial measures: profit, revenues, and costs. The measures satisfy the following business model relation: profit = revenues – costs (or in shorthand notation $y = x_1 - x_2$) and are associated with the dimensions Time (t), Location (l), and Product (p). Time, Location, and Product dimensions follow the hierarchies: Quarter $\prec$ Year $\prec$ All-Times, City $\prec$ All-Locations, and ProdName $\prec$ All-Products. Suppose an analyst is exploring the cube at the Time $\times$ Location $\times$ Product plane as shown in Table 1. The analyst notices a significant increase in profit in the year 2002 compared to the norm year 2001. A significant increase or decrease in a variable is called a symptom.

Product (p)	P1			
Time (t)	Location (l)	Profit (y)	Revenues (x_1)	Costs (x_2)
2001	A	15	35	20
	B	10	50	40
	Location average	12.5	42.5	40
	All-Locations	**25**	**85**	**60**
2002	A	10	40	30
	B	20	60	40
	Location average	15	50	35
	All-Locations	**30**	**100**	**70**
All-Times	A	25	75	50
	B	30	110	80
	Location average	27.5	92.5	65
	All-Locations	**55**	**185**	**130**

Table 1: Part of the multidimensional financial data for the ABC-company (2-D representation). The two boxes ('All-locations', year 2001 and 2002) with dark boundaries indicate the two values being compared.

With the existing tools the analysts has to find the reason(s) for this drop by manually drilling down the numerous different planes underneath it, inspecting the entries for big drops and drilling down further. This process can get rather problematic especially for typically larger real-life datasets. We propose to use an explanation model for finding the answer to this

question. The user, for example, simply highlights the two cells and invokes the "explanation" operator. The results as shown in Table 2 are presented as a list of *one-level* explanations that give the causes for the symptom "increase in profit". The explanations are one level deep, in the sense that they are based on only one relation from the dimension hierarchy (e.g. Quarter $\prec$ Year and City $\prec$ All-Locations or measures (profit = revenues – costs). In the table the first two rows show the causes for the symptom based on the Location dimension. For Location "A" the influence-value is negative so this results in a *counteracting cause*, and for Location "B" the influence-value is positive so this results in *contributing cause*. The influence-value indicates what the quantitative difference between the actual (e.g. the year 2002) and norm value (e.g. the year 2001) of a variable (e.g. profit) would have been if only one independent variable (e.g. revenues, profit in location "A", or profit in quarter "q1") or would have deviated from its norm value. Contributing causes increase the probability of the effect and counteracting causes decrease the probability of the effect. The next four rows account for the difference by explanation in the hierarchy of the Time dimension. These rows show that the third and fourth quarter are responsible for the profit increase in the year. The influence-value for the second quarter is 0, so this variable does not contribute to the explanation of the observed symptom. The last two rows explain the difference following the relation between the measures; namely the business relation profit = revenues – costs. The increase in the measure profit is due to the increase in the measure revenues despite the fact that the costs have risen for All-Locations in the year 2002. Detailed definitions for contributing causes, counteracting causes, influence measures will be given in section 4.

Product	P1			
Time	Location	Measure	Cause	Influence-value
All	A	Profit	Counteracting	-5
All	B	Profit	Contributing	10
2002.q1	All	Profit	Counteracting	-1
2002.q2	All	Profit	No	0
2002.q3	All	Profit	Contributing	2
2002.q4	All	Profit	Contributing	4
2002	All	Revenues	Contributing	15
2002	All	Costs	Counteracting	-10

Table 2: One-level deep explanations for the increase in the measure Profit marked in Table 1.

Furthermore, explanation generation can be continued automatically into the direction of the dimension hierarchy and into direction of the measures for the identified symptom. Based on Table 1 we can proceed with explanation generation, for example, for the contributing cause the revenues in 2002 for All-Products (row 7). This cause can be considered a lower level symptom and is explained further by using the dimension hierarchies for the dimension Time and Location or some business model relation like Revenues = Volume · Unit Price. The method can filter out insignificant influences by defining so-called parsimonious sets [3, 4].

2. Overview of the multidimensional model

2.1. Multidimensional data

We now shortly review the basic principles of multidimensional data. Each measure m_i can be analyzed using a set of dimensions $\{d_1, d_2, \ldots, d_n\}$. The measures that can be analyzed by the same set of dimensions are described by the *base cube*. A base cube uses level instances

of the lowest dimension levels of each of its dimensions to identify a measure value. The relationship between a set of measure values and the set of identifying level instances is called a *cell*. Loading data into the OLAP data cube means that new cells will be added to the base cubes, whereas also new level instances may be added to dimension levels. If a dimension d_k is related to a measure m_i by means of the base cube, then the dimension hierarchy of d can be used to aggregate the measures values of m using operators like SUM, COUNT, and, AVG. We assume a many-to-one relationship between the level instances of two dimension levels $d_k[p]$, $d_k[q]$ (if we refer to level p of dimension d_k, we write $d_k[p]$) with dimension hierarchy $d_k[p] \prec d_k[q]$ to ensure correct aggregation of measure values. Most multidimensional data models require that the dimension hierarchy of a dimension is strict [1, 13, 14]. Aggregating measure values along the hierarchies of different dimensions (i.e. rollup) creates a multidimensional view on data, which is known as the data cube or cube. This type of organisation provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data. Typical OLAP operations are: *rollup* (aggregation on a data cube), *drilldown* (reverse of roll-up), *slice* (selection on one dimension) and *dice* (defines a sub-cube) and *pivot* (rotates the data axes).

The dimension hierarchy (or classification structure) has a *schema component* and an *instances* (values) *component*. That is the dimension levels and their structure constitute the schema, and the dimension level instances constitute the instances (data) for this schema. Shoshani et al. [12] developed a graph model that separates the dimension levels and the level instances into two representation forms. For example, the dimension hierarchy of the time dimension of the last example is represented at the meta-data level (intentional representation) as shown in Figure 1 at the left. Underlying this representation the system stores and maintains the instance and their relationship, called the extensional representation (the right side of the figure).

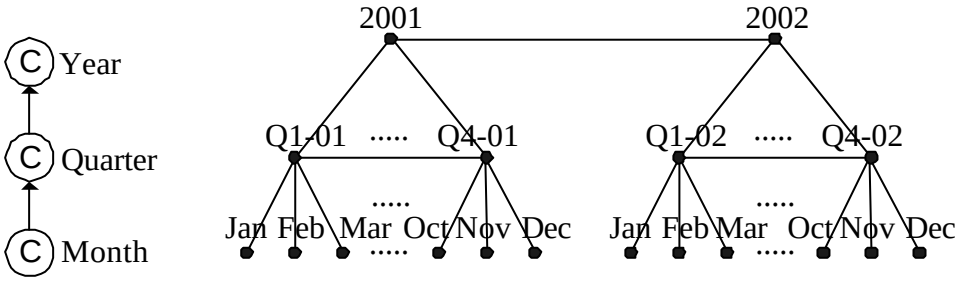

Figure 1: Intensional representation and extensional representation

2.2. Summarizability in the multidimensional model

An important criterion for the quality of OLAP cube design is the correctness of aggregations. The summarizability of OLAP databases is an important property because violating this condition can lead to erroneous conclusions and decisions. In [8] Lenz and Shoshani have studied summarizability in OLAP and statistical databases. We briefly mention the three necessary conditions for summarizability: *disjointness* of dimension levels in dimension hierarchies, *completeness* in dimension hierarchies, and *compatibility* of measure attributes types (*flow*, *stock*, and *value-per-unit*) with statistical functions.

3. Overview of the explanation model

3.1 A causal model of explanation and diagnosis

According to a causal model of explanation, phenomena (events) are explained by giving their causes. In this research paper the exposition on diagnostic reasoning and causal explanation is largely based on Feelders and Daniels' notion of explanations [3, 4], which is essentially based on Humpreys' notion of aleatory explanations [7] and the theory of explaining differences by Hesslow [6]. Causal influences can appear in two forms: contributing and counteracting. Therefore, Humphreys proposes the following canonical form for causal explanations:

$$\text{Event } E \text{ occurred because of } C^+, \text{ despite } C^-,$$

where E is the event to be explained, C^+ is non-empty set of contributing causes, and C^- a (possibly empty) set of counteracting causes. The explanation itself consists of the causes to which C^+ jointly refers. C^- is not part of the explanation of E, but gives a clearer notion of how the members of C^+ actually brought about E.

The explanandum introduced by Feelders and Daniels is a three-place relation $\langle a, F, R\rangle$ between an object a (e.g. the ABC-company), a property F (e.g. having a low profit) and a reference class R (e.g. other companies in the same branch or industry). Here the event E is thus replaced by a more detailed explanandum. The task is not to explain why a has property F, but rather to explain why a has property F *when the members of R do not*. For the purpose of explanation, the class R can often be reduced to one member r, which is in some sense the average of the class R or the ideal object. The syntax of an explanation reads:

$$\langle a, F, r\rangle \text{ because } C^+, \text{ despite } C^-.$$

3.2. The business model

Feelders states that explanations are usually based on general laws expressing relations between events, such as cause effect relations or constraints between variables. These laws are represented in a business model M. The model M, which is a form of domain knowledge, can be derived from many domains, like finance, accounting, logistics, and so forth. The business model M represents quantitative variables by means of mathematical equations of the form:

$$y = f(\mathbf{x}) \text{ where } \mathbf{x} = (x_1, \ldots, x_n).$$

In Table 3, an example is given of a business model. The business model M is associated with a directed graph $E(M)$, called the *explanatory graph*. The explanatory graph of the business model in Table 3 is depicted in Figure 2.

1. Gross Profit = Revenues - Cost of Goods
2. Revenues = Volume · Unit Price
3. Cost of Goods = Variable Cost + Indirect Cost
4. Variable Cost = Volume · Unit Cost
5. Indirect Cost = 30% · Variable Cost

Table 3: Example business model M

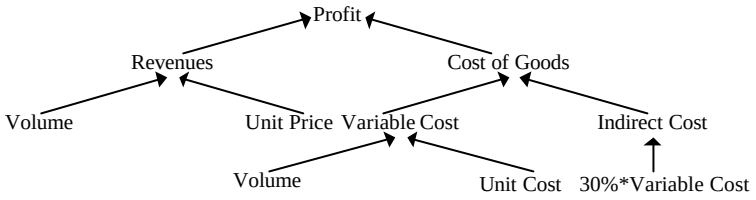

Figure 2: Explanatory graph for business model *M*

The form of the business model relations is such that exactly one variable appears on the left hand side of the relation. Economists often specify their models in such a way that this requirement is met.

3.3. The norm model

The norm model specifies which reference object(s) should be used to compare. It also specifies the variables with respect to which the comparison should be made. In [4], the most common "reference objects" to diagnose business performance are described:

- Theoretical norm values
- Historical norm values
- Industry (or company) averages as norm values
- Plans and budgets as norm values

Through theoretical norm values, one tries to establish a norm for a particular financial, accounting, or operating variable that is applicable to all companies or business units. A historical norm value for a particular variable is its value in one or more previous time periods. The industry average of companies operating within the same industry, or the company average of comparable business units is often used as a norm for the company or business unit. For a particular company, the norm values may be the result of an explicit planning process. A plan may for example indicate the production to be achieved or it may contain budget values for particular expense items.

4. Diagnosis and explanation in OLAP

4.1. Diagnosis and explanation for multidimensional data

In this section we build on the theory and methodology for automated diagnosis as described in [3, 4]. The concepts of this methodology are adapted to use them for automated diagnosis and explanation on multidimensional data. In order to apply the explanandum on multidimensional data we have to link the explanation model with the multidimensional model. For this purpose the attributes of multidimensional data - measures and dimensions - have to be connected with the elements of the explanandum.

To make the connection for the dimension attribute we have to define the actual object *a* and the reference object *r* as multidimensional objects with, for example, a time, location, or product dimension. Therefore, we associate the object *a* and the reference object *r* of the explanandum with the *dimension* vector **d**. The vector $\mathbf{d} = (d_1, d_2, \ldots, d_n)$ denotes the *n*-component dimension vector. In addition, the property *F* is related to the *measure* attribute of the multidimensional model. The actual object *a* corresponds to an aggregate (or subcube) in the data cube that needs to be explained. The reference object *r* expresses the aggregate used for reference in the data cube. Both the actual object *a* and reference object *r* are formed by

aggregating one ore more dimensions. We are interested in explaining the difference between aggregate *a* and *r*. Consequently we have to explain the following type of events in the data cube:

- *a* = the actual multidimensional object, e.g. sales(2003,All-Products);
- *F* = a particular measure deviates from its norm value, e.g. having a decrease in sales;
- *r* = the multidimensional reference object, e.g. sales(2002,All-Products).

Because the multidimensional actual and reference objects will be clear by the selection of the analysts, we can now simplify the explanation format to:

$$\partial y = q \text{ occurred because } C^+, \text{ despite } C^-.$$

In this expression, $\partial y = q$ specifies an event in the data cube, i.e. the occurrence of a qualitative difference between the *actual* and the *norm* value of y, denoted by y^a and y^r, respectively. The actual and norm value represent cells for comparison in the data cube. In comparison, the actual object a and reference object r denote two subcubes formed by expanding the common aggregated dimensions of y^a and y^r. The qualitative difference can take on one of the values {low, normal, high}.

A diagnosis is an explanation for observed abnormal behaviour of a variable, sometimes called *problem identification*. Problem identification is a process that computes a value $g(y^a, y^r)$ for each variable, where g is some user-specified function such as percentage difference or absolute difference. If this value is below (above) some specified threshold, a symptom $\partial y = low$ (or $\partial y = high$) is added to the list of symptoms. The result of problem identification is a set of symptoms $S = \{\partial y_1 = q_1, \ldots, \partial y_n = q_n\}$ where $q_i = \{low, high\}$. The next two paragraphs discuss the knowledge representation structures for diagnosis in the multidimensional data.

4.2. Knowledge representation structures in multidimensional data

In multidimensional data there is structure in the dimensions – the dimension hierarchy – and in the measures – the business model. Both structures can be used as "explanation directions". Explanation generation in the dimensions is directed towards lower dimension levels of the dimension hierarchy; it uses the aggregation relation (drilldown equation) between the parent and child dimension level. We investigate the common situation where the aggregation relation is the summarization of measures in the dimension hierarchy. Moreover, explanation generation in the measures is directed towards the right-hand side of the business model equations. Business model equations represent relations between measures. The business model (1) and drilldown equations (2) have subsequently the following general forms:

(1) $y = f(\mathbf{x})$, and
(2) $d_k[p] \prec d_k[q]$.

The measure vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ specifies the business model variables, where each variable is associated with the dimensions of the multidimensional model. $d_k[p] \prec d_k[q]$ is some part of the dimension hierarchy with dimension levels *p* and *q*. The dimension level *p* is

defined as $p = q + 1$. The two explanation directions are illustrated by the following instances for equations (1) and (2) derived from the multidimensional dataset for the ABC-company:

(ins. 1) profit(All-Periods,B,P1) = revenues(All-Periods,B,P1) − costs(All-Periods,B,P1),
(ins. 2) profit(All-Periods,B,P1) = profit(2001,B,P1) + profit(2002,B,P1)

where (ins. 1) is a business model equation, and (ins. 2) is a drilldown equation for the time dimension with the dimension hierarchy $t[\text{Year}] \prec t[\text{All-Times}]$.

We elaborate on the common situation where the form of the aggregation function in the dimension hierarchy of a dimension is additive. Therefore additive relations exist between the dimension levels of a dimension hierarchy. These additive functions are now defined formally. If the measure y is *additive* the following top-down definition can be derived:

Definition 1 Measure y is additive in dimension d_k with dimension hierarchy $d_k[p].\mathbf{j} \prec d_k[q].i$, with dimension levels *p* and *q*, and instances *i* and *j* if:

$$y(\ldots, d_k[q].i, \ldots) = \sum_{j=1}^{n} y(\ldots, d_k[p].j, \ldots).$$

The dimension d_k is an arbitrary dimension out of the dimension vector $\mathbf{d}$. Where *i* is the parent instance (e.g. the year "2002") on dimension level *q* (e.g. *t*[Year]), and the children in vector **j** are the instances (e.g. the quarters "q1", "q2", "q3", and "q4" of 2002) of dimension level *p* (e.g. *t*[Quarter]). The number of elements on the right-hand side of the equation is equal to *n*; the number of level instances in dimension level *p* associated to *one* level instance *a* of dimension level *q*. The summarization equation in *definition 1* is called a *drill-down* equation.

Definition 2 Measure *y* is additive if *Definition 1* holds for all dimensions in $\mathbf{d}$ and all levels of the dimension hierarchies.

For example, profit, revenues and costs are additive measures in Table 1, according to Definition 1 and 2. These measures are typical "flow measures", and can therefore always be correctly aggregated. In contract, the measures unit sale price and unit costs in Table 3 are non-additive because they are of the type "value-per-unit". This type of measure does not have the summarizability property in the dimensions and dimension levels.

The length of the equation in Definition 1 depends on the structure of the level instance component of the data. Therefore, the length of the equation has to be determined. The number of level instances *n*, defined as the number of level instances on level *p* associated with one particular level instance of level *q*, can be determined by using a COUNT operator. This operator simply counts the number of instances of a particular dimension level.

4.3. The norm model in multidimensional data

The actual variable for comparison y^a, can in theory be compared with every other cell in the data cube, the reference variable y^r. However, in general only the cells on the same aggregation levels will be used as norm values for obvious reasons (like the measurement scale of the variable). For example, in Figure 1 the actual value for comparison is the cell

y^a(2002,A,P1) and y^r(2001,A,P1) denotes the reference object. In this example, the two cells only differ in only one dimension. We can equally well handle cases where the two cells differ in more than one dimension as long as we have common dimensions on which both the cells are aggregated.

In addition, special type of norm values can be added to the data cube, namely *norm values based on the average* ($\overline{y}^r$) of each dimension (level). These type of norm values can be calculated for each individual dimension (level) in **d**. This idea will be illustrated by extending the example of the ABC-company with a larger financial dataset, with data for multiple years (1999-2004), locations (A-F), and products (P1-P4). Figure 3, shows the data cube for this example with the measure variable profit (*y*) on the dimensions Time[Year], Location [City], and Product[ProdName]. The actual cell y^a(1999,A,P1) can be compared with the average profit of: the time dimension ($\overline{y}^r$(All-Periods,A,P1)), the location dimension ($\overline{y}^r$(1999,All-Locations,P1)), and the product dimension ($\overline{y}^r$(1999,A, All-Products).

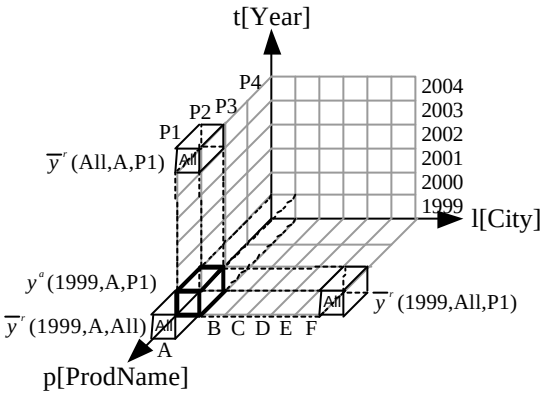

Figure 3: Norm values based on average in a cube for the ABC-Company

To be more formal about this idea, we now introduce some notation. If the measure y is additive, then the norm values y^r are also additive. Therefore summarization relations exists between successive dimension levels in a dimension for all norm values y^r. In particular, a norm value for a particular dimension d_k can be summarized up to a higher level on the dimension hierarchy. A norm value y^r on dimension vector **d** with a dimension hierarchy $d_k[p] \prec d_k[q]$, with levels p and q and instances i and j has the following summarization relation when *y* is a an additive measure:

$$y^r(\ldots, d_k[q].i, \ldots) = \sum_{j=1}^{n} y^r(\ldots, d_k[p].j, \ldots).$$

The relation between reference objects based on averages in the multidimensional model is defined analogously: $\overline{y}^r(\ldots, d_k[q].i, \ldots) = (\sum_{j=1}^{n} y^r(\ldots, d_k[p].j, \ldots))/n$.

4.4. Diagnosis and explanation in OLAP

For diagnosis and explanation in the multidimensional model we use the methodology as described in [3, 4]. We apply the proposed definitions for the *measure of influence*,

contributing and counteracting causes, and *parsimonious sets*. First we discuss the situation where explanation is sustained by the business model $y = f(\mathbf{x})$ first, after that we elaborate on the situation where explanation is sustained by drilldown equations.

To determine the contributing and counteracting causes that explain the difference between the actual and the norm value of y, a measure of influence for the business model is defined as follows:

$$\inf(x_i, y) = f(\mathbf{x}^r_{-i}, x^a_i) - y^r,$$

where $f(\mathbf{x}^r_{-i}, x^a_i)$ denotes the value of $f(\mathbf{x})$ with all variables and associated dimensions evaluated at their norm values, except x_i. And where *i* is the index of the vector $\mathbf{x}$. In words, $\inf(x_i, y)$ indicates what the difference between the actual and norm value of y would have been if only x_i would have deviated from its norm value. The correct interpretation of the measure of influence depends on the form of the function *f*; the function *f* has to satisfy the so-called *conjunctiveness constraint*. This constraint captures the intuitive notion that the influence of a single variable should not turn around when it is considered in conjunction with the influence of other variables. Two types of functions satisfy this constraint, namely monotonic and additive functions. For an elaboration on restrictions on the equations we refer to [3, 4].

For the drilldown equations the measure of influence can be simplified because the function *f* is additive by definition. Therefore, the influence measure can always be correctly interpreted. The measure of influence for explanation generation in a dimension hierarchy is defined as:

$$\inf(y(\ldots, d_k[p].j, \ldots), y(\ldots, d_k[q])i, \ldots) = y^a(\ldots, d_k[p].j, \ldots) - y^r(\ldots, d_k[p].j, \ldots).$$

In the definition d_k is one dimension out of the vector **d**, the other dimensions (and possible associated dimension hierarchies) remain constant in the calculation of the influence measure. In words, $\inf(\ldots, y(d_k[p].j, \ldots), y(\ldots, d_k[q]).i, \ldots)$ gives the difference between the actual and norm values of the parent *i* if only the single child *j* would have deviated from its norm value.

Contributing and counteracting causes for explanations in the business model follow the proposed definitions. The set of contributing (counteracting) causes C^+ (C^-) consists of components x_i of $\mathbf{x}$ out of the business model with: $\inf(x_i, y) \times \Delta y > 0$ (< 0). In words, the contributing causes are those variables whose influence values have the same sign as Δy, and the counteracting causes are those variables whose influence values have the opposite sign. The contributing and counteracting causes for explanations in the dimension hierarchy of a dimension can be defined similarly. Now the set of contributing (counteracting) causes C^+ (C^-) consists of the set of instances on dimension level *p* out of the dimension hierarchy of dimension d_k with $\inf(y(\ldots, d_k[p].j, \ldots), y(\ldots, d_k[q].i, \ldots)) \times \Delta y > 0$ (< 0).

Parsimonious sets of causes are used as a filter measure to leave insignificant influences out of the explanation to prevent an information overload of the analyst. The same filter is applied in explanation generation for multidimensional data. The parsimonious set of contributing causes is the smallest subset of the set of contributing causes, such that its influence on *y* exceeds a particular fraction (T^+) of the influence of the complete set. The

fraction T^+ (T^-) will typically be close to one. In the sequel we use the following format for parsimonious one-level explanations:

$$\partial y = q \text{ occurred because } C_p^+, \text{ despite } C_p^-.$$

5. Maximal explanation in the multidimensional model

5.1. Introduction and background

In this section we present an explanation generation algorithm for symptoms discovered in multidimensional data. The algorithms for maximal explanation in multidimensional data clearly built on the idea of a "maximal explanation tree" as described in [3, 4]. However, we extend the basic idea of generating explanations in multiple equations of business model equations to maximal explanation in the measures and dimensions of multidimensional data. We start by giving some background information about the analyses of multidimensional data.

In [5], Han remarks that the data mining process at multiple dimension levels may proceed in several ways: *progressive deepening*, *progressive generalization*, and *interactive up-and-down*. The explanation generation process for multidimensional data is, in this respect, quite similar to the knowledge mining process at multiple dimension levels. Especially, the idea of progressive deepening seems very "natural" in the explanation generation process; find an explanation on a high level in the dimension hierarchy and progressively deepen it to find the explanations for events at lower levels of the dimension hierarchy.

The idea of progressive deepening resembles the strategy of analysts in analyzing multidimensional data. When an analyst queries the multidimensional data to make a decision, he usually follows an *incremental top-down* approach in creating and analyzing cubes [8, 17]. First, the analyst creates a very "coarse-grained" cube that describes the variables for which a decision should be made. Second, in carrying out multidimensional analyses the analyst compares, aggregates, and transforms, etc. the cells of this cube. Finally, the analyst creates a more detailed cube for the level instances for which further analysis is necessary.

5.2. Maximal explanation in multidimensional data

The idea of progressive deepening is used in the construction of a maximal explanation algorithm for multidimensional data. Such an algorithm is needed to create multi-level explanations. Thus far, we have only discussed "one-level" explanations. The explanations are one-level deep, in the sense that they are based on a single relation from the business model or on a single drilldown relation derived from the dimension hierarchy of a dimension. For diagnostic purposes, however, it is useful to continue an explanation of $\partial y = q$ where $q = \{\text{low}, \text{high}\}$, by explaining the qualitative differences between the actual and norm values of its contributing causes. This process can be continued until a parsimonious contributing cause is encountered that cannot be explained:

- within the business model, because the business model equations do not contain a relation in which this contributing cause appears on the left-hand side, and
- within the dimensions, because the drilldown equations do not contain a summarization relation in which this contributing cause appears on the left-hand side.

The result of this process is a *maximal explanation tree of causes*, where *y* is the root of the tree with two types of children, corresponding to its parsimonious contributing and counteracting causes respectively. A node that corresponds to a parsimonious contributing cause is a new symptom that can be explained further. And a node that corresponds to a parsimonious counteracting cause has no successors.

Two special types of explanation trees can be constructed, namely an explanation tree based on only business model equations (1), and an explanation tree based on only drilldown equations from dimension d_k (2). In (1) explanation generation continuous until a variable cannot be explained any further in the business model. Here all derived explanations are on the same aggregation level and in each step a new measure will be encountered as a cause. In (2) the dimension d_k is associated with a hierarchy $d_k[z] \prec d_k[z-1] \prec \ldots \prec d_k[0]$. Where dimension level $d_k[0]$ or $d_k[\text{All}]$ is the highest dimension level and $d_k[z]$ is the lowest dimension level. This hierarchy can be translated in the following set of drilldown equations:

$$y(\ldots, d_k[q].i, \ldots) = \sum_{j=1}^{n} y(\ldots, d_k[p].j, \ldots),$$

where $q = 0,1,\ldots,z-1$, and $p = q+1$. Explanation generation continues until the lowest level of the dimension hierarchy; the level $d_k[z]$. In this tree the derived explanations are related to one measure, and in each step the aggregation level is lowered by one level.

In the maximal tree of causes explanation generation may continue in the direction of the business model or in direction of the drilldown equations of the dimensions. In other words, in some nodes of the tree on can alternate the use of drilldown and business model equations. In this way different explanation paths exist from the root of the explanation tree to the endnotes.

There is a unique canonical way to form reference objects in the next explanation step from the previous step. For each derived contributing cause we can construct a reference object based on the right hand sides of the business model (1) and drilldown equations (2), if there is an equation that sustains further explanation:

(1) $y^r = f(\mathbf{x}^r)$,

(2) $y^r(\ldots, d_k[q].i, \ldots) = \sum_{j=1}^{n} y^r(\ldots, d_k[p].j, \ldots)$.

In this way a chain or tree of reference objects is constructed. In fact this is the total reference subcube or aggregate *r*. As said, a very common reference object is the reference object based on some average. Also for this type of reference object the canonical way of forming reference objects holds. However the averages have to be computed and added to the subcube *r* by taking:

(1) $\bar{y}^r = f(\mathbf{x}^r)$ for the business model equations, and

(2) $\bar{y}^r(\ldots, d_k[q].i, \ldots) = (\sum_{j=1}^{n} y^r(\ldots, d_k[p].j, \ldots))/n$ for the drilldown equations.

We now propose an algorithm to produce a maximal explanation tree of causes for symptoms in multidimensional data. The algorithm uses all the business model equations and drilldown equations together with the canonical reference object. We define a recursive procedure for maximal explanation of multidimensional data:

Algorithm 1 (**Maximal explanation of multidimensional data**). *A maximal explanation of multidimensional data* $\partial y(\mathbf{d}) = q$ *is a tree with the following properties*:

1. *Construct a business model sub-tree with explanation of the business model equations.*
2. *For all nodes of the business model sub-tree construct dimension sub-trees for the dimensions in* $\mathbf{d}$ *with explanation of the drilldown equations.*
3. *For each node where one of the dimensions is not on the lowest level of the dimension hierarchy, or where one of the measures has a business model equation, do Maximal explanation of multidimensional data. (Algorithm 1).*

6. Related work

To position this paper we now discuss some related work regarding the explanation of differences and the exploration of multidimensional data. We will explain these previous initiatives in terms of the general explanation formalism as formulated in paragraph 4.1.

Sarawagi [10] presented an operator for OLAP data cubes that lets the analyst get summarized reasons for drops or increases observed at an aggregated level. This operator eliminates the need to manually drill-down for such reasons. Sarawagi developed an information theoretic formulation for expressing these reasons and designed a dynamic programming algorithm for it. In terms of the explanation model Sarawagi compares the actual value in subcube C_b (= object a) with the expected one according to subcube C_a (= reference object r) and the compact summary of the difference table A (= property F) for each elementary cell. A consists of rows not only detailed but also aggregated levels of the cube. With each row of A a *ratio r* - y^a / y^r - is associated that indicates to the user that everything underneath that row had the same ratio. The ratio of a cell results from the ratio of immediate predecessor cell and known ratios of neighbouring cells. The idea is to find A such that a user reconstructing C_b from C_a and A will incur the smallest amount of error. This can be achieved by listing rows that are significantly different than their parents (parsimonious counteracting causes) and aggregating rows that are similar (parsimonious contributing causes) such that the error due to summarization is minimized. The error is calculated, based on a probability distribution (e.g., the normal distribution, Poisson distribution) around the expected value: $\Pr(C_b \mid C_a, A)$. The goal of the sender is the deviation of A such that the total error is minimized.

Sarawagi et al. [11] developed a discovery-driven exploration paradigm that mines the data for exceptions and summarizes the exceptions at appropriate levels in advance. The discovery-driven method is guided by pre-computed indicators of exceptions at various levels of detail in the cube. The model they use is inspired by the table analysis method used in statistical literature. In [11] a value in a cell of a data cube is an exception (or symptom) if it is significantly different from the expected value based on a statistical model. This model computes the expected value of a cell $\hat{y}$ in context of its position in the data cube and combines trends along different dimensions that the cell belongs to. The difference between the actual value y (= object a) and the expected value $\hat{y}$ (= reference object r)) determines the "degree of surprise" a cell. However the deviation of the actual value from the expected value must be larger that some threshold after normalization value to call it an exception. The property F of the explanation model holds if $|y - \hat{y}| / s > q$. For a value y in a cube, $\hat{y}$ is defined as a function f of contributions from various higher level group-bys as:
$\hat{y} = f_{aggregation}(g)$. The g terms are the coefficients of the model equation. The function f can take one an additive and multiplicative forms. The coefficients are estimated from the base data by mean-based or median-based (trimmed-mean) estimates. By this method the user is

guided by the model to interesting data regions using pre-computed indicators. In comparison with the explanation formalism, this model does not generate parsimonious contributing or counteracting causes, but is more a model to identify symptoms (outliers) automatically.

7. Case study: sport equipment sales data

We use a demo dataset (called "GOSales") obtained from the Cognos OLAP product PowerPlay [2] as a case study for the algorithm. The dataset has 42.063 records and four dimensions (see the star schema in Figure 4). The Vendor dimension is not used for explanation generation because of space limitations. The fact table "Financial Facts" present the measures of the dataset. The numbers within brackets denote the cardinality of that level.

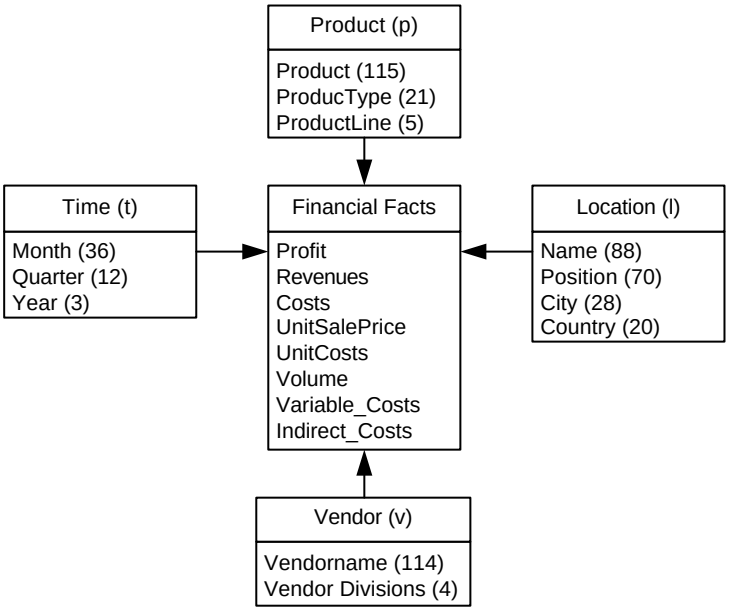

Figure 4: Star schema describing the dimensions and measures of the GOSales dataset

The relations between the measures - the business model - are shown in Table 3. An important condition for summarizability is compatibility of the measures with the statistical aggregation function applied. Two types of measures are present, namely: "flow" (Profit, Revenues, Costs, Variable Cost, Indirect Cost, and Volume) and "value-per-unit" (Unit Sale Price and Unit Cost). The flow measures are summarized, however the summarization of value-per-unit measures is not meaningful semantically [5]. Therefore, we apply the weighted average on these measures. In calculating the average of the value-per-unit measures we take into account the volumes associated with it. Most measures in the business model are additive except the measures Unit Sale Price and Unit Cost.

We now provide an example of a complete diagnosis in the business model and in the dimensions of the GoSales dataset. The diagnosis will be performed for the Netherlands ("NL") in the year 2001, which has a decrease in profit compared to the profit in the year 2000. Problem identification yields the set of symptoms $S = \{\partial\text{profit}(2001,\text{NL},\text{All})=\text{"}low\text{"}\}$ since the relative difference between norm value and actual value, $(353{,}096.80 - 515{,}715.03)/515{,}715.03 = -0.32$, is below some specified lowbound of -0.10. A full specification of the event to be explained in the cube is:

$$< \text{profit}^a(2001,\text{NL},\text{All}), \partial profit = \text{"low"}, \text{profit}^r(2000,\text{NL},\text{All}) >.$$

We want to explain this event in the direction of the business model as well as in the direction of the Time, Location, and Product dimension. We want to omit insignificant influences from the explanations, therefore we take $T^+ = T^- = 0.7$. Firstly, we start with explanation in the direction of the business model. Hence the corresponding equation in Table 3 is: $\text{profit}(t, l, p) = \text{revenues}(t, l, p) - \text{costs}(t, l, p)$. Therefore, $\text{profit}^a(2001,\text{NL},\text{All})$, or in short $\text{profit}^a(2001,.,.)$, is the root of the explanation tree. Computation of the influences of the individual variables in the business model equation for profit yields the following results and calculations:

	Norm	Actual	inf
profit(.,.,.)	515,715.03	353,096.80	
revenues(.,.,.)	4,087,870.34	4,341,657.58	253,787.24
costs(.,.,.)	3,572,155.31	3,988,560.78	-416,405.47

Table 4: Data for explanation of $\partial\text{profit}(2001,\text{NL},\text{All}) = "low"$

From the data in Table 4 the following one-level explanation is obtained: $\partial\text{profit}(2001,\text{NL},\text{All}) = "low"$, because $C_p^+ = \{\text{costs}(.,.,.)\}$, despite $C_p^- = \{\text{revenues}(.,.,.)\}$. Or in words the profit has gone down in the Netherlands in the year 2001 because the costs have risen despite the fact that the revenues have gone up. Explanation generation continues for the contributing cause, therefore, the event to explained is specified as $<\text{costs}^a(2001,.,.)$, ∂costs =”high”, $\text{costs}^r(2000,.,.)>$. Table 5 summarizes the model results for the high costs.

	Norm	Actual	inf
costs(.,.,.)	3,572,155.31	3,988,560.78	
variable_costs(.,.,.)	2,747,811.78	3,068,123.68	320,311.90
indirect_costs(.,.,.)	824,343.53	920,437.10	96,093.57

Table 5: Data for explanation of $\partial\text{costs}(2001,\text{NL},\text{All}) = "high"$

From the data in the table is follows that $C_p^+ = \{\text{variable\_costs}(.,.,.)\}$ and $C_p^- = \{\ \}$. The model filters out the indirect_costs(.,.,.) from the contributing causes. The reason is that its contribution to the overall contributing influence $(\text{inf}(C^+, \text{costs}(.,.,.)) = 416405.47$ on total costs is negligible. The increase in the variable costs can be explained further in the business model. Now the event to be explained is $<\text{var\_costs}^a(2001,.,.)$, $\partial\,\text{var\_costs}$ =”high”, $\text{var\_costs}^r(2000,.,.)>$. The explanation is sustained by equation 4 from Table 3.

	Norm	Actual	inf
variable_costs(.,.,.)	2,747,811.78	3,068,123.68	
avg_unit_costs(.,.,.)	56.42092276	58.487224084	100,633.01
volume(.,.,.)	48702	52458	211,916.99

Table 6: Data for explanation of $\partial\text{variable\_costs}(2001,\text{NL},\text{All}) = "high"$

From Table 6 it can be concluded that $C_p^+ = \{\text{avg\_unit\_cost}(.,.,.), \text{volume}(.,.,.)\}$, since both the (weighted) average unit costs and the sales volume contributed to the difference between norm value and actual value, and both are needed to explain the desired fraction. Obviously, $C_p^- = \{\ \}$.

The previous examples of one-level diagnosis can be combined to a complete diagnosis in the business model. Figure 5, summarizes the results of the maximal diagnosis in the direction of the measures for event $\partial\text{profit}(2001,NL,All) = \text{"}low\text{"}$, where dotted lines indicate counteracting causes.

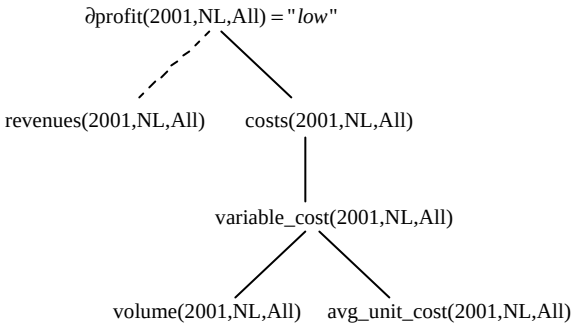

Figure 5: Diagnosis for symptom $S = \{\text{profit}^a(2001,NL,All) = \text{"}low\text{"}\}$ in the business model

Secondly, we continue with explanation in the direction of the dimensions for this particular event. Therefore, the algorithm uses the drilldown equations of the dimensions Time, Location, and Product. We start with explanation in the direction of the Time dimension. This dimension is associated with the dimension hierarchy Time[Month] $\prec$ Time[Quarter] $\prec$ Time[Year]. Because *profit* is an additive cube function the dimension hierarchy can be translated into the following drilldown equation: $\text{profit}(2001,NL,All) = \sum_{i=1}^{4} \text{profit}(\text{[quarter]}.a_i,NL,All)$. The norm values for the individual quarters are determined by taking the canonical reference object.

	Norm	Actual	inf
profit(.,.,.)	515,715.03	353,096.80	
profit(*.q1,.,.)	18,038.73	40,168.30	22,129.57
profit(*.q2,.,.)	249,197.75	105,965.76	-143,231.99
profit(*.q3,.,.)	106,681.04	90,982.64	-15,690.40
profit(*.q4,.,.)	141,797.50	115,980.10	-25,817.40

Table 7: Data for explanation of $\partial profit(2001,NL,All) = \text{"}low\text{"}$

In Table 7 comparison is made between the quarters of the year 2000 (norm) with the quarters of the year 2001. From the data in Table 8 it follows that $C_p^+ = \{\text{profit}(2001.q2,.,.)\}$, because only the second quarter is needed to explain the desired fraction of $(\text{inf}(C^+,\text{profit}(2001,.,.))$. Obviously, $C_p^- = \{\text{profit}(2001.q1,.,.)\}$. The contributing cause is explained further, because the drilldown equation: $\text{profit}(2001.q2,NL,All) = \sum_{i=1}^{3} \text{profit}(\text{[month]}.a_i,NL,All)$ sustains this ex-planation in the dimension hierarchy. Now the event to be explained is $<\text{profit}^a(2001.q2,.,.)$, $\partial\text{profit} = \text{"low"}$, $\text{profit}^r(2000.q2,.,.)>$. Without giving the data and calculation we give the causes: $C_p^+ = \{\text{profit}(2001.Apr,.,.),\text{profit}(2001.May,.,.),\text{profit}(2001.Jun,.,.)\}$ despite $C_p^- = \{\ \}$. Now explanation stops in the Time dimension because there is no dimension level below the month level in the hierarchy. Figure 6 under the node *time*, summarizes the results of the diagnostic process in the dimension hierarchy of the Time dimension.

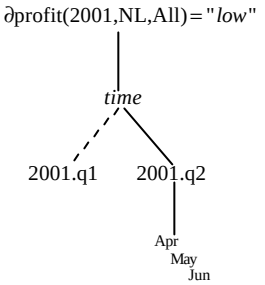

Figure 6: Diagnosis for symptom $S = \{\text{profit}^a(2001,\text{NL},\text{All}) = \text{"}low\text{"}\}$ in the Time dimension

In summary, Figure 5 and 6 present examples of special type of explanation trees where explanation is sustained by respectively: only business model equations, and drilldown equations. The figures only summarize step 1 and part of step 2 of the algorithm for maximal explanation of multidimensional data. Step 2 of the algorithm continues with diagnosis in the Location and Product dimensions. In step 3 explanation generation proceeds for contributing causes that have an equation that supports further explanation. This final step is omitted in this case study.

8. Summary and conclusions

In this paper, we presented a formal framework for explanation and diagnosis in multidimensional data. For the construction of explanations in an OLAP system the elements of the explanation formalism are adapted to the structure of multidimensional data. Therefore, explanation generation in an OLAP system proceeds in two directions: in the direction of the business model equations (the measures) and in the direction of the drilldown equations (the dimension hierarchies). The algorithm as proposed uses the concept of a maximal explanation tree of causes, where explanation generation is continued until a parsimonious contributing cause cannot be explained anymore by a drilldown or business model equation. In addition, the algorithm uses a unique canonical way to form reference objects in successive explanation steps. The result of the algorithm is a large semantic tree which can be presented to the analyst. We demonstrated the algorithm by applying it on a demo multidimensional dataset with dimension hierarchies and a financial business model.

We believe that this framework could assist analysts in generating explanations for symptoms in multidimensional data. Moreover, the framework can easily be applied to all kinds of financial or accounting models. In general, the novel framework could lead to better decisions based on multidimensional business data, especially when the dataset is large. The result of this research can be used to develop an analytical tool as an add-in for OLAP systems.

References

[1] L. Cabibbo, R. Torlone: *"A logical approach to multidimensional databases"*, in EDBT'98, Proc. of 6th intl. conf. on extending database technology, Valencia, Spain, pages 183-197, Springer LNCS 1377, March 23-27, (1998).

[2] Cognos Incorporated, Cognos Series 7, Cognos PowerPlay, http://www.cognos.com, (2004).

[3] H.A.M. Daniels, A.J. Feelders, *"Theory and methodology: a general model for automated business diagnosis"*, European Journal of Operational Research, 130: 623-637, (2001).

[4] A.J. Feelders, *"Diagnostic reasoning and explanation in financial models of the firm"*, PhD thesis, University of Tilburg, (1993).

[5] J. Han: *"Mining knowledge at multiple concept levels"*, In Proc. 4th Int'l Conf. on Information and Knowledge Management (CIKM'95), Baltimore, Maryland, pp. 19-24, Nov., (1995).

[6] G. Hesslow: *"Explaining differences and weighting causes"*, Theoria, 49:87-111, (1984).

[7] P.W. Humphreys: *"The chances of explanation"*, Princeton University Press, Princeton, New Jersey, (1989).

[8] H.J. Lenz, A. Shoshani, *"Summarizability in OLAP and statistical databases"*, 9th International Conference on Statistical and Scientific Database Management (SSDBM), (1997).

[9] OLAP Council, http://www.olapcounsil.org, (2000).

[10] S. Sarawagi: "Explaining differences in multidimensional aggregates", Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, pages 42-53, (1999).

[11] S. Sarawagi, R. Agrawal, and Nimrod Megiddo: *"Discovery-driven exploration of OLAP data cubes"*, In Proc. of the 6th Int'l Conference on Extending Database technology (EDBT), Valencia, Spain, (1998).

[12] A. Shoshani, *"OLAP and statistical databases: Similarities and differences"*, In Sixteenth ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems, pages 185-196, (1997).

[13] T. Thalhammer, M. Schrefl, M. Hohania: *"Active data warehouses: complementing OLAP with analysis rules"*, Data & Knowledge engineering, 39, 241-269, (2001).

[14] P. Vassiliadis, T. Sellis: *"Modeling multidimensional databases, cubes, and cube operations"*, in M. Rafanelli and M. Jarke, editors, 10th intl. conf. on scientific and statistical database management, proceedings, Capri, Italy, pages 53-62, IEEE Computer Society Press, July 1-3, (1998).