Automated detection and explanation of exceptional values in a datamining environment

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

In this paper, we describe an extension of the datamining framework with automated causal diagnosis, offering the possibility to automatically detect and explain exceptional values to support business decision tasks. This functionality can be built into the conventional OLAP (On-Line Analytical Processing) or datamining system using a generic explanation formalism, which mimics the work of business decision makers in diagnostic processes. The diagnostic process is now carried out manually by (business) analysts, where the analyst explores the multidimensional data to spot exceptions, and navigate the data to find the reasons for these exceptions. Such functionality can be provided by extending the conventional datamining system with an explanation formalism, which mimics the work of human decision makers in diagnostic processes. Here diagnosis is defined as finding the best explanation of unexpected behaviour (symptoms or exceptional values) of a system under study. This definition assumes that we know which behaviour we may expect from a correctly working system, otherwise we would not be able to determine whether the actual behaviour is what we expect it or not. The expected behaviour in a datamining environment can be derived from some statistical model or can be expert knowledge from analysts. The central goal is the identification of specific knowledge structures and reasoning methods required to construct computerized explanations from multidimensional data and business models. A methodology that automatically generates explanations for exceptional values in multidimensional business data is proposed. The methodology was tested on a case study involving the comparison of financial results of a firm's business units.

Keywords: Datamining, Multidimensional databases, OLAP, Explanation, Business Intelligence.

1. Introduction

Today's OLAP (On-Line Analytical Processing) and datamining systems have limited explanation or diagnosis capabilities. The diagnostic process is now carried out manually by (business) analysts, where the analyst explores the multidimensional data to spot exceptions, and navigate the data with operators like drill-down, roll-up, and selection to find the reasons for these exceptions. Such functionality can be provided by extending the conventional datamining system with an explanation formalism, which mimics the work of human decision makers in diagnostic processes. Here diagnosis is defined as finding the best explanation of unexpected behaviour (symptoms or exceptional values) of a system under study [18]. This definition assumes that we know which behaviour we may expect from a correctly working system, otherwise we would not be able to determine whether the actual behaviour is what we expect it or not. The expected behaviour or norm model in a datamining environment can be derived from some statistical model or can be expert knowledge from analysts.

The core component of a datamining 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. A popular data model is the multidimensional or OLAP database, also known as the *data cube*, where data is organized using the dimensional modelling approach, which classifies data into *measures* and *dimensions*. Measures like, for example, sales figures and costs, are the basic units of interest for analysis. Dimensions correspond to different perspectives for viewing measures. Dimensions are usually organised as *dimension hierarchies*, which offer the possibility to view measures at different dimension levels (e.g. $month \prec quarter \prec year$). Some typical OLAP operations for interactive querying and analysis are: *rollup* (i.e. aggregation on a data cube), *drilldown* (i.e. reverse of roll-up), *slice* (i.e. selection on one dimension) *dice* (i.e. defining a sub-cube), and *pivot* (i.e. rotates the data axes).

The objective of this paper is to extend the OLAP system with a complete diagnostic process. Two important phases in the diagnostic process are: *problem identification* and *explanation generation*. In the problem identification phase the OLAP data is mined for exceptional values. And when a discrepancy between actual and norm behaviour is discovered, and is qualified as unacceptable with respect to some specified norm, the next step is to explain this discrepancy using our "understanding" of the multidimensional model. In short, we automate the current user-driven analysis of OLAP data, with an explanation formalism that finds exceptions, and subsequently finds out why exceptions have emerged.

Our exposition on diagnostic reasoning and causal explanation is largely based on Feelders and Daniels' notion of explanations in [3, 5], which is essentially based on Humpreys' notion of aleatory explanations [11] and the theory of explaining differences by Hesslow [7]. The canonical form for causal explanations is taken from [3, 5]:

$$\langle a, F, r\rangle \text{ because } C^{+}, \text{ despite } C^{-}. \tag{1}$$

where $\langle a, F, r\rangle$ is the symptom 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, but gives a clearer notion of how the members of C^{+} actually brought about the symptom. The explanandum is a three-place relation 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). 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*. This general formalism for explanation constitutes the basis of the framework for diagnosis in a OLAP/datamining context developed in this paper.

To position this paper we discuss some related work regarding the explanation of differences and the exploration of multidimensional data. In [13] Sarawagi presented an operator for data cubes that lets the analyst get summarized reasons for drops or increases observed at an aggregated level. In [14] the authors 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. In comparison with the explanation formalism, this model does not generate causes, but is a model to identify symptoms.

The remainder of this paper is organized as follows. Section 2 introduces the notation for the multidimensional model, followed by a description of normative models appropriate for diagnosis in Section 3. In Section 4 the explanation formalism is extended for multidimensional data in order to automatically generate explanations, and in Section 5 the complete method is illustrated in a case study on OLAP sales data. Finally, conclusions are discussed in Section 6.

2. Notation and equations

Many different notations and definitions of multidimensional data schemata can be found in literature [1, 16, 17]. Here we introduce a generic notation that is particular suitable for combining the concepts of measures, dimensions, and dimension hierarchies. A measure is defined as a function on multiple domains: $y^{i_1 i_2 \ldots i_n} : D_1^{i_1} \times D_2^{i_2} \times \ldots \times D_n^{i_n} \to \mathbb{R}$. Each domain D_i has a number of hierarchies ordered by $D_i^{i_{\max}} \prec D_i^{i_{\max}-1} \prec \ldots \prec D_i^0$, where D_i^0 is the highest level and $D_i^{i_{\max}}$ is the lowest level in D_i. A dimension's top level has a single level instance $D_i^0 = \{\text{All}\}$. For example, for the time dimension we could have the following hierarchy $\text{T}^2 \prec \text{T}^1 \prec \text{T}^0$, where $\text{T}^0 = \{\text{All-T}\}$, $\text{T}^1 = \{2000{,}2001\}$, and $\text{T}^2 = \{\text{Q1,Q2,Q3,Q4}\}$. Sometimes we will write T[Quarter] for T^2. A *cell* is denoted by $(d_1, d_2, \ldots, d_n)$, where d_i are elements of the domain hierarchy at some level, so for example (2000,Amsterdam,Beer) is a cell. Each cell contains data, which are the values of the measures y like, for example, sales^{122}(2000, Amsterdam,Beer). If no confusion can arise we will leave out the upper indices indicating level hierarchies and write sales(2000,Amsterdam,Beer). Furthermore, the combination of a cell and a measure we call a *data point*. It is important to mention that a cell can be present at multiple aggregation levels or *contexts*. For example, suppose that (2001.Q2,Germany) is a cell in a 2-dimensional data cube with the dimensions Time and Location and the hierarchy $quarter \prec year$. The possible contexts for the above cell are the following: {(Year.Q2,Germany), (2001.Quarter,Germany), (2001.Q2,Country), (Year.Quarter,Germany), (2001.Quarter,Country), (Year.Q2,Country), and (Year.Quarter, Country)}.

The measure values at the lowest level possible ($D_i^{i_{\max}}$) are entries of the *base cube*. If a measure value is on the base cube level, then the hierarchies of the domains can be used to aggregate the measure values using aggregation operators like SUM, COUNT, or, AVG. By applying suitable equations, we can alter the level of detail and map low level cubes to high level cubes and vice versa. For example, aggregating measure values along the dimension hierarchy (i.e. rollup) creates a multidimensional view on the data, and de-aggregating the measures on the data cube to a lower dimension level (i.e. drilldown), creates a more specific cube. Here we investigate the common situation where the aggregation operator is the summarization of measures in the dimension hierarchy. So *y* is an *additive measure* if in each dimension and hierarchy level of the data cube:

$$y^{i_1 \cdots i_{q-1} \cdots i_n}(\ldots, a, \ldots) = \sum_{j=1}^{J} y^{i_1 \cdots i_q \cdots i_n}(\ldots, a_j, \ldots)\,, \tag{2}$$

where $a \in D_i^{q-1}$, $a_j \in D_i^q$, q is some level in the dimension hierarchy, and *J* represents the number of level instances in D_i^q. An example equation corresponding to two roll-up operations reads:

$$\text{sales}^{102}(2001,\text{All-Locations},\text{Beer}) = \sum_{j=1}^{4}\sum_{k=1}^{20} \text{sales}^{212}(2001.\text{Q}_j,\text{Country}_k,\text{Beer})\,.$$

If there is no confusion about the level in the dimension hierarchy we will use *y*(…,+,…), for the left-hand side of the above equation. In general, a "+" in place of an instance denotes that summing has occurred in that domain. In this way a data cube with only two dimensions is represented by a table where the row totals are given by $y(d_1,+)$, column totals are given by $y(+,d_2)$, and the grand total is given by $y(+,+)$.

Furthermore, we assume a business model M is given representing relations between measures. These relations can be derived from many domains, like finance, accounting, logistics, and so forth. Relations are denoted by

$$y^{i_1 i_2 \ldots i_n}(d_1, d_2, \ldots, d_n) = f(\mathbf{x}^{i_1 i_2 \ldots i_n}(d_1, d_2, \ldots, d_n)), \tag{3}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, and $y, x_1, \ldots, x_n$ are measures defined on the same domains. Business model equations mostly hold on equal aggregation levels in the data cube, therefore we may leave out upper indices if no confusion can arise. In Table 1, an example of a business model with quantitative relations from a sales database is given.

Table 1
Example business model *M*

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

3 Normative models

The normative model specifies the reference class that should be used to compare. It also specifies the measures with respect to which the comparison should be made. We distinguish between two broad classes of reference objects namely: *external* and *internal*. External reference objects refer to norm values that are derived from other sources then the data under consideration, and internal reference objects are based on data in the database. Examples of external norm values are industry averages or plans and budgets [5].

There are many ways to construct internal reference objects for multidimensional data. The simplest way is pairwise comparison [13], where a value of a measure *y* is compared with another in the data cube, the reference variable norm(*y*) In general, only the cells on the same aggregation levels will be used for obvious reasons, like the measurement scale of the variable. For example, we can compare sales(2000,Germany,All-Products) with the sales of the previous year, norm(sales(1999,Germany,All-Products)), as an historical norm value.

Other common internal norm values are the average or expected value of a cell. We use the following notation:

$$\bar{y}(\ldots,+,\ldots) = \frac{1}{J}\sum_{j=1}^{J} y(\ldots, a_j, \ldots), \tag{4}$$

and for the average over all domains we write $\bar{y}(+,+,\ldots,+)$. Expected values are based on statistical models. A huge variety of statistical models exists for two-way (contingency) tables, three-way tables, etc., see Scheffé [15] and Tukey in [8, 9]. Here we only consider two models namely the additive model and the model of independence. For a multidimensional data set, in the situation of only two dimensions, we can write the expected value as an additive function of three terms obtained from the possible aggregates of the table:

$$\hat{y}(d_1, d_2) = \bar{y}(d_1, +) + \bar{y}(+, d_2) - \bar{y}(+,+). \tag{5}$$

Where we assume that the joint contribution of the aggregates is the sum of the separate contributions from each aggregate and $e(d_1,d_2) \sim N(0,s^2)$. The coefficients of the model are estimated by OLS.

Another standard model to compute expected values logically follows from the independence assumption. Consider a two-dimensional cube (two-way table) with row variables (in D_1) and column variables (in D_2) where the values in the cells represent counts of some measure *y*. The null hypothesis is that there is no association between the row variable and the column variable. We can model the data as a multinomial distribution with row and column variables. Now, let $p(d_1,d_2)$ denote the unknown *probability* of an observation being in the cell (d_1,d_2) of the table, and $p(d_1,+)$, $p(+,d_2)$, are the marginal probabilities of the row and column variables. Then fom the multiplication law of probability, independence between the row and the column variables implies that:

$$p(d_1,d_2) = p(d_1,+)\,p(+,d_2) \qquad (6)$$

The probabilities are estimated from the realisations in the table, it can be shown that $\hat{p}(d_1,+) = y(d_1,+)/y(+,+)$, and $\hat{p}(+,d_2) = y(+,d_2)/y(+,+)$ (these are also the maximum likelihood estimates). Now the estimate of the "expected value" under the null hypothesis of independence is given by:

$$\hat{y}(d_1,d_2) = \frac{y(d_1,+)\,y(+,d_2)}{y(+,+)}\,. \qquad (7)$$

In addition, the null hypothesis can be tested with Pearson's chi-squared statistic. Similar definitions can be found for multidimensional tables in literature on contingency tables [5].

Now the expression $\partial y = y - \text{norm}(y) = q$ ($q \in \{\text{low,high}\}$) specifies a symptom in the data cube, i.e. the occurrence of a quantitative difference between the actual and norm value. Problem identification is the process that computes a value $g(y,\text{norm}(y))$ for each cell, where *g* is some user-specified function such as percentage difference or absolute difference. We can scale the residual with the standard deviation $s(y)$. In that case, a cell is a exceptional value or surprise value [14] if $(y - norm(y))/s$ is higher than some threshold d.

4. Methodology

4.1. Explanation in multidimensional databases

Explanations of events are usually based on general laws expressing relations between events, such as cause effect relations. In the data cube, two types of relations are available for explanation generation namely: multiple additive relations in the dimension hierarchies (a) and the business model relations between measures (b) like, for example:

(a) $\text{profit}^{102}(\text{2000,Amsterdam,Beer}) = \sum_{j=1}^{4} \text{profit}^{202}(2000.\mathit{quarter}_j\text{, Amsterdam,Beer})$,

(b) $\text{profit}^{102} = \text{revenues}^{102} - \text{costs}^{102}$.

We are interested in explaining the difference between object *a* and *r*. Contributing and counteracting causes that explain ∂y are determined by the calculation of a *measure of influence* [5, 10]. The correct interpretation of the measure depends on the form of the function *f*; the function 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. The conjunctive-ness constraint holds for monotone and additive functions [3, 5], which frequently occur in the business model and the hierarchical dimension relations.

First we discuss the situation where explanation is supported by relations in the business model, after that we elaborate on the situation where explanation is supported by dimension hierarchies. The definition of the measure of influence for a model equation reads:

$$\inf(x_i, y) = f(\mathrm{norm}(\mathbf{x}_{-i}), x_i) - \mathrm{norm}(y), \tag{8}$$

where $f(\mathrm{norm}(\mathbf{x}_{-i}), x_i)$ denotes the value of $f(\mathbf{x})$ with all measures evaluated at their norm values, except the measure x_i. All measures are evaluated at the same aggregation level. In words, $\inf(x_i, y)$ indicates what the difference between the actual and norm value of y would have been if only the measure x_i would have deviated from its norm value. In the dimension hierarchy f is additive by definition, it follows from (7) that:

$$\inf(y^{i_1\ldots i_q\ldots i_n}(\ldots,a_j,\ldots), y^{i_1\ldots i_{q-1}\ldots i_n}(\ldots,a,\ldots)) = y^{i_1\ldots i_q\ldots i_n}(\ldots,a_j,\ldots) - \mathrm{norm}(y^{i_1\ldots i_q\ldots i_n}(\ldots,a_j,\ldots)). \tag{9}$$

When explanation is supported by a business model equation the set of contributing (counteracting) causes C^+ (C^-) consists of measures x_i 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 ∂y, and the counteracting causes are those variables whose influence values have the opposite sign. If explanation is supported by the dimension hierarchy, the set of contributing (counteracting) causes C^+ (C^-) consists of the set of child instances a_j of dimension level i_q out of the hierarchy of a specific dimension with $\inf(y^{i_1\ldots i_q\ldots i_n}(\ldots,a_j,\ldots),$ $y^{i_1\ldots i_{q-1}\ldots i_n}(\ldots, a,\ldots)) \times \Delta y > 0 \ (< 0)$.

4.2. Reducing the number of explanations

Because every applicable equation yields a possible explanation, the number of generated explanations for a single symptom can be quite large. Especially when explanations are chained together to form a tree of explanations we might get lost in an intractable branching tree. In order to leave insignificant influences out of the explanation we introduce three generic concepts.

Firstly, in the problem identification phase the analyst distillates a set of symptoms. This means that if a cell does not have a large deviating value – based on some statistical model or defined by a user – it is not identified as a symptom and therefore not considered for explanation generation.

Secondly, small influences are left out in the explanation by a filter. In [3, 5] the set of causes is reduced to the so-called *parsimonious set of causes*. The parsimonious set of contributing causes C_p^+ 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^+ is a number between 0 and 1, and will typically 0.8 or so.

A third way to reduce the number of explanations is by applying a *measure of specificity* for each applicable equation. This measure quantifies the “interestingness” of the explanation step. The measure is defined as:

$$\text{specificity} = \frac{\text{\# possible causes}}{\text{\# actual causes}}. \tag{10}$$

The number of possible causes is the number of right-hand side elements of each equation, and the number of actual causes is the number of elements in the parsimonious set of causes. Using this measure of specificity we can order the explanation paths from specific to general and if desired only list the most specific steps.

4.3. Multi-level explanations

The explanation generation process for multidimensional data is quite similar to the knowledge mining process at multiple dimension levels. Especially, the idea of *progressive deepening* [6] seems very "natural" in the explanation generation process; start symptom detection on an aggregated level in the data cube and progressively deepen it to find the causes for that symptom at lower levels of the dimension hierarchy or business model. This idea we will adopt for so-called *multi-level explanations*. In the previous parts, we have discussed "one-level" explanations; explanations based on a single relation from the business model or dimension hierarchy. For diagnostic purposes, however, it is meaningful to continue an explanation of $\partial y = q$, by explaining the quantitative differences between the actual and norm values of its contributing causes. In multi-level explanation this process is continued until a parsimonious contributing cause is encountered that cannot be explained further because:

- the business model equations do not contain an equation in which the contributing cause appears on the left-hand side.
- the dimension hierarchies do not contain a hierarchical equation in which the contributing cause appears on the left-hand side.

The result of this process is an *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. In the explanation tree there are numerous explanation paths from the root to the leaf nodes. This implies that many different explanations can be generated for a symptom. In most practical cases one would therefore apply the pruning methods discussed above yielding a comprehensive tree of the most important causes. Interchanging the order of equations in explanation generation may give different explanatory trees, in general commutativety does not always hold. In fact commutativety only holds if after expanding the original equation by substitution, the same end-result is obtained for both orders.

4.4. Canonical chain of reference objects

The explanation model applies an equation of the form (2) or (3) to generate causes for an identified symptom. In order to identify causes, reference objects have to be formed in the explanation generation process for all the RHS variables of the equation used for explanation. Now we state that there is a natural way to construct reference variables for variables on the RHS of the equation. The basic idea is that *the context and reference object selected for determining the reference value of the LHS variable are the basis for determining the reference objects for the RHS variables* of the equation used for explanation. In addition, if the RHS variable has a relation in which it appears on the LHS the construction of reference can be continued for the RHS variables of this relation following the same basic idea. In this way, a *chain of reference objects* is formed for the next explanation step from the previous step. Although there is a canonical way in forming reference objects in the explanation generation process typical situations exist in forming reference objects for the RHS variables. The use of a drilldown equation of the form (2), or the use of a business model equation of the form (3) enforces different requirements on the chain of reference objects.

In the next two paragraphs we elaborate on these typical situations and present examples to illustrate them. In the presented examples we suppose that first a symptom (∂*profit*= "high" or "low"; from now we write *y* for the measure *profit*) is detected using a particular reference object y^r in a context $(D_1^{q_1},\ldots,D_i^{q_i-1},\ldots,D_n^{q_n})$ of the GoSales data cube [2] using the (simple additive) multi-way ANOVA model. Now the next step in the diagnostic process is to explain the symptom using the explanation model and canonical reference objects for the RHS variables in each explanation step.

The typical situation in constructing reference objects for RHS variables $x_1, x_2, \ldots, x_n$ of a business model equation is that the *same context and statistical model* are applied as in determining the reference object for the LHS variable *y*. It has to be remarked that the context and statistical model have to be related to the RHS variable (measure) x_i under consideration. We illustrate this with the following example.

Example 1

Suppose that we have detected the symptom $\partial \mathit{profit}(\text{2001,Germany})=\text{"low"}$ in the context (Year,Country). To derive this symptom we computed the additive model $\hat{y}$(2001,Germany) $= \bar{y}(2001,+) + \bar{y}(+,\text{Germany}) - \bar{y}(+,+)$ to get the reference value for the actual datapoint $y^a(\text{2001,Germany})$. We now apply the business model relation $\text{profit}^{110} = \text{revenues}^{110} - \text{costs}^{110}$ (in short $y = x_1 - x_2$) to generate explanations. Therefore, canonical reference values for the measures x_1 and x_2 have to be determined in the (same) context $x_i(\text{Year,Country})$ with the additive model $\hat{x}_i\,(\text{2001,Germany}) = \bar{x}_i\,(2001,+) + \bar{x}_i\,(+,\text{Germany})$ $-\bar{x}_i\,(+,+)$ where $i = \{1,2\}$.

The typical situation in constructing reference objects for RHS variables of drilldown equations is that reference objects have to be determined in the context $y(D_1^{q_1},\ldots,D_i^{q_i},\ldots,D_n^{q_n})$, derived with a drilldown on dimension $D_i^{q_i-1}$ (the dimension selected for explanation), with application of the same statistical model. The obvious remark is that the context and statistical model have to be related to the actual RHS variables $y^{i_1\ldots i_q\ldots i_n}(\ldots,a_j,\ldots)$ under consideration. We illustrate this with the following example.

Example 2

Now we explain the symptom of Example 1 in the dimension hierarchy of the Location dimension. We apply the drilldown equation $y^{110}(\text{2001,Germany}) = \sum_{j=1}^{2} y^{120}(\text{2001,Germany.}$ $\text{City}_j)$ for explanation generation. The reference object for the LHS variable y^{110}(2001,Germany) was determined in the context (Year,Country), therefore the canonical reference objects for the RHS variables have to be determined in the context (Year,Country.City). To derive this context we drill-down in the dimension hierarchy of the Location (*L*) dimension from the level Country (L^1) to the level City (L^2). Furthermore, because Year and Country are on the rolled-up level in the initial context for determining the reference object for the LHS variable, they remain on the roll-up level in determining the reference objects for the RHS variables. The reference values are computed with the additive model $\hat{y}(2001, \text{Germany.City}_j) = \bar{y}^{120}(2001,+) + \bar{y}^{120}(+,\text{Germany.City}_j) - \bar{y}^{120}(+,+)$.

5. Case study: sport equipment sales data

We use a dataset (called "GOSales") obtained from the Cognos OLAP product PowerPlay [2] as a case study for our method. The data consist of 42.063 records and four dimensions organised in a star schema; see Fig. 1.

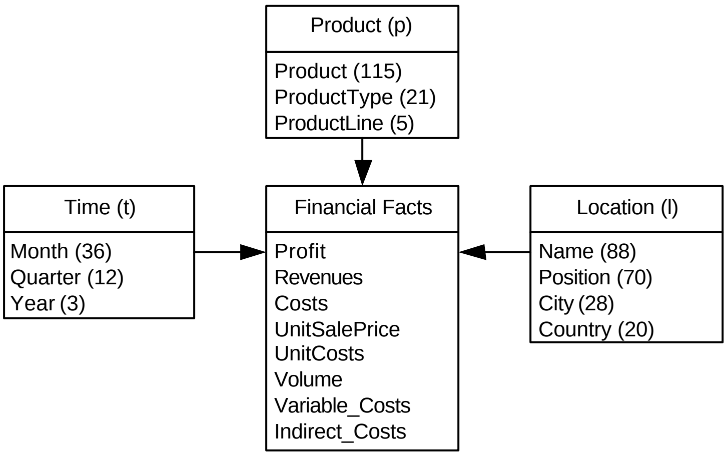

Fig. 1. Star schema describing the dimensions and measures of the GOSales dataset

In the fact table "Financial Facts" the measures of the dataset are listed. The numbers within brackets denote the cardinality of that level. The relations between the measures are defined in Table 1. An important condition for summarizability is compatibility of the measures with the statistical function applied. Two types of measures are present, namely: "flow type" (Profit, Revenues, Costs, Variable Cost, Indirect Cost, and Volume) and "value-per-unit type" (Unit Sale Price and Unit Cost). The flow measures are summarized however the summarization of value-per-unit measures is not meaningful [12]. Therefore, we take weighted averages of these measures. In calculating the average we take into account the volumes associated with it.

In the case study we present examples of the explanation process namely: the multi-level explanation of a symptom in the product dimension, the special handling of norm values in the time dimension, and explanation in the business model *M* of the data cube. The explanation formalism is implemented in MS Excel, and the diagnosis was carried out using this prototype. Symptom detection in the cube starts on an aggregated level, where the analyst has to select the context, the combination of aggregation levels from the domains, from where to start the explanation generation process. Suppose that an analyst starts exploring the cube in the context (Year,Country,All-Products) and problem identification yields the symptom S= {∂profit(2001,Spain,All-Products)="*low*"}. Under the assumptions of the additive model we calculate the expected values for the context, using (5): $\hat{y}(\text{year,country}) = \bar{y}(\text{year},+) + \bar{y}(+,\text{country}) - \bar{y}(+,+)$, and standardize the residuals. Here we choose $d = 1.28$ corresponding to a probability of 90% in the normal distribution, and find that the standardized residual for Spain in the year 2001 (=1.33) is larger than the threshold value. A full specification of the event to be explained is: < y (2001,Spain,All-Products), ∂profit="low", norm($\hat{y}$ (2001,Spain, All-Products))>. We want to omit insignificant influences from the explanations, therefore we take $T^{+} = T^{-} = 0.75$. Explanation generation may start in the product dimension (P) for the detected symptom, where explanation is sustained by additive relations. First the decrease in profit on the All-Products level is examined on the ProductLine level of the dimension hierarchy P[Product] $\prec$ P[ProductType] $\prec$ P[ProductLine] $\prec$ P[All-Products]. Hence the first corresponding additive equation using (2) applied for explanation generation is:

$$\text{profit}^{110}(.,.,\text{All-Products}) = \sum_{j=1}^{5} \text{profit}^{111}(.,.,\text{ProductLine}_j)$$

Therefore, $\text{profit}^{110}(2001,\text{Spain},\text{All-Products})$ is the root of the explanation tree. The norm values for explanation generation are based on the expected values for the entries of the dimension level ProductLine in the context (Year,Country,ProductLine). Computation of the influences of the individual variables for the additive equation above with (9) yields the results in Table 2. From the data in Table 2 it can be concluded that $C_p^+ =$ {profit(.,.,Personal Accessories), profit(.,.,Golf Equipment), profit(.,.,Mountaineering Equipment)}, since only these three relatively large causes are needed to explain the desired fraction of $\text{inf}(C^+,\text{profit}(.,.,\text{All-products})$. Obviously, $C_p^- = \{\ \}$.

Table 2
Data for explanation of S = {∂profit(2001,Spain,All-Products) = ”low”}

Profit(2001,Spain,All-Products)	Norm	Actual	Inf
Profit(.,.,All-Products)	242,169.03	145,976.67	
Profit(.,.,Camping Equipment)	6,488.07	-8,684.36	-15.172.43
Profit(.,.,Personal Accessories)	46,610.41	22,521.12	-24,089.29
Profit(.,.,Outdoor Protection)	17,807.01	10,033.18	-7,773.83
Profit(.,.,Golf Equipment)	99,048.87	79,928.64	-19,120.23
Profit(.,.,Mountaineering Equipment)	72,214.67	42,178.09	-30,036.58

The parsimonious causes are explained further on the level ProductType, the data for comparison of the entries on the level ProductType of the ProductLine Personal Accessories is presented in Table 3. From the data in Table 3 it follows that $C_p^+ =$ {profit(.,.,Watches),profit(.,.,Knives),profit(.,.,Binoculars),profit(.,.,Navigation)}, and $C_p^- = \{\ \}$. Here the relatively large cause Watches contributes significantly to the low profit for Personal Accessories, therefore explanation generation continues downwards in the Product level. However the other parsimonious causes (Knives, Binoculars, and Navigation) are presented on the aggregated level, to avoid too much detail for the analyst.

Table 3
Data for explanation of S = {∂profit(2001,Spain,Personal Accessories) = ”low”}

Profit(2001,Spain,Personal Accessories)	Norm	Actual	Inf
Profit(.,.,Personal Accessories)	46,610.41	22,521.12	
Profit(.,.,Personal Accessories.Watches)	24,409.07	13,345.47	-11,063.60
Profit(.,.,Personal Accessories.Eyewear)	4,686.22	2,302.91	-2,383.31
Profit(.,.,Personal Accessories.Knives)	6,664.91	2,535.36	-4,129.55
Profit(.,.,Personal Accessories.Binoculars)	4,792.63	1,385.56	-3,407.07
Profit(.,.,Personal Accessories.Navigation)	6,057.59	2,951.82	-3,105.77

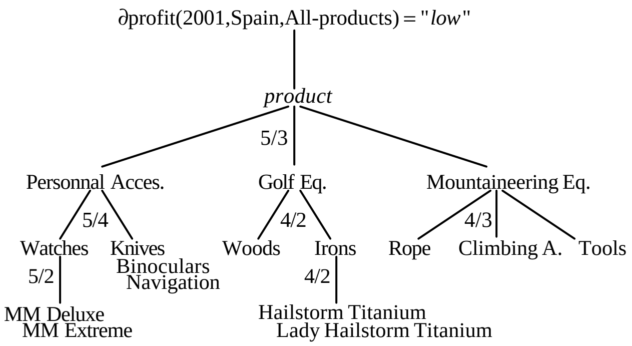

Fig. 2. Diagnosis S = {∂profit(2001,Spain,All-Products)="low"} in the product dimension

Now the previous examples of one-level explanations are combined to a complete diagnosis in the product dimension. Fig. 2 summarizes the results of the multi-level diagnosis, where the lines indicate contributing causes (possible dotted lines indicate counteracting causes) and the numbers indicate the specificity value of the explanation step. The specificity values are determined using (10). Moreover, explanation trees can be constructed in the same way for the time and location dimension.

In addition, the symptom S={∂profit(2001,Spain,All-Products)="*low*"} can also be explained in the business model *M* of the data cube. Hence the corresponding equation in Table 1 is: $\text{profit}^{110} = \text{revenues}^{110} - \text{costs}^{110}$. The norm values for the measures revenues(.,.,.) and costs (.,.,.) are both based on the context (Year,Country,All-Products). Computation of the influences of the individual measures in the equation for profit by applying (8) yields the following results:

Table 4
Data for explanation of S = {∂profit(2001,Spain,All-Products) = "low"}

(2001,Spain,All-Products)	Norm	Actual	Inf
Profit(.,.,.)	242,169.02	145,976.67	
Revenues(.,.,.)	2,269,708.27	1,698,895.10	-570,813.17
Costs(.,.,.)	2,027,539.25	1,552,918.43	474,620.82

From the data in Table 4 it follows that: profit(.,.,.)="*low*", because $C_p^+ = \{\text{revenues(.,.,.)}\}$, despite $C_p^- = \{\text{costs(.,.,.)}\}$. Explanation generation may continue for the contributing cause. The results of Table 4 for the one-level diagnosis in the business model are summarized in Fig. 3.

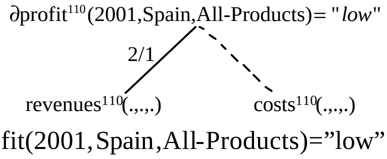

Fig. 3. Diagnosis S = {∂profit(2001,Spain,All-Products)="low"} in the business model *M*

6. Summary and conclusion

In this paper, we presented a formal framework for explanation and diagnosis in datamining systems, in particular, for the multidimensional or OLAP model. Explanation generation in multidimensional data can proceed in two directions: in the direction of the business model equations and downwards in the dimension hierarchies. The methodology as proposed uses the concept of an explanation tree of causes, where explanation generation is continued until a parsimonious contributing cause cannot be explained further. The result of the process is a semantic tree, where the main causes for a symptom can be presented to the analyst. To prevent information overload to the analyst, several techniques are discussed to prune the explanation tree. The methodology is demonstrated by applying it on a multidimensional sales dataset with dimension hierarchies and a financial business model.

We believe that this framework could assist analysts in generating explanations for exceptional values in multidimensional data. Moreover, the framework can easily be applied to all kinds of business 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-on 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] B.S. Everitt: "The analysis of contingency tables", Chapman & Hall, (1977).

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

[6] 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).

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

[8] D.C. Hoaglin, F. Mosteller, and J.W. Tukey: "Exploring data tables, trends and shapes", Wiley series in probability, (1988).

[9] D.C. Hoaglin, F. Mosteller, and J.W. Tukey: "Understanding Robust and Exploratory Data Analysis", John Wiley, New York, (1983).

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

[11] D.W. Kosy, B.P. Wise: *"Self-explanatory financial planning models"*, in Proceedings of AAAI-84, Los Altos, CA, Morgan Kaufmann, pages 176-181, (1984).

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

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

[14] S. Sarawagi, R. Agrawal, and Nimrod Megiddo: *"Discovery-driven exploration of OLAP data cubes"*, Proc. of the 6th Conference on EDBT, Valencia, Spain, (1998).

[15] H. Scheffé: "The analysis of variance", New York, Wiley, (1959).

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

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

[18] W. Verkooijen: *"Automated financial diagnosis: a comparison with other diagnostic domains"*, Journal of Information Science, 19, pages 125-135, (1993).