Business Intelligence
SS 2018

Modeling in BI

W. Grossmann

Contents

- Models an Modeling in BI
- Logical and Algebraic Structures
- Graph Structures
- Analytical Structures
- Models and Data

Models and Modeling in BI

- There are many different models used in BI
  - Examples you know:

Models and Modeling in BI – Definition of a Model

- Model Definition: Models represent some part of the business process and allow precise formulation of interesting questions (Analytical Goals)
  - How can we realize the representation? (***representation function***)
  - How should we formulate the representation? (“***model language***”)

Models and Modeling in BI – Representation Function of Models

- Models of Phenomena
  - Phenomena: Features of the business process interesting from an analytical point of view
  - Models define a picture of the phenomena (caricatures)
    - Idealized models, e.g., control flow of the business process, a treatment process, a course design
    - Analogical models: Overtake ideas from other sciences, e.g., gravity model for relations between persons in dependence of distance
    - Phenomenological models: Statistics, e.g. regression

Models and Modeling in BI – Representation Function of Models

- Models of Data
  - We have no precise idea about the models, but only a number of candidate models for the empirical data
  - The task is to learn the most appropriate model (Machine Learning, Data Mining)
  - Simple example: Churn management:
    - Which variables influence the churn behavior of a customer , e.g., age, sex, marital status, income, ….?
    - How should we define the relation between churn behavior and theses variables?

Models and Modeling in BI – Representation Function of Models

- Models of Theories
  - Each application domain of BI has specific domain knowledge, usually defined by concepts and relation (logical relations) between the concepts
  - Concepts and logical relations define a formal system (ontology)
  - Understanding this formal system as a theory data instances are models of this theory

→ Database models

Models and Modeling in BI – Languages for Models

- Corresponding to the multitude of models there are different formulations (languages) used:
  - UML or ER-modeling for data
  - BPMN for formulation of the control flow
  - Statistics in case of modeling customer behavior
  - Connectedness (reachability) in a graph

Models and Modeling in BI – Formulation of Models

- Each language has its own semantic allowing definition of certain model elements an formulation of generic questions
  – Queries in a database
  – Simultaneous occurrence of two events in a business process
  – Strength of association between two variables
  – Graph models for social networks

Models and Modeling in BI – Formulation of Models

- Generic questions can be formulated in different languages
  - Example: Relations between attributes
    - Formulate a query in a data model and represent the result as a table
    - Define a regression model and formulate the relation as an equation
    - Use a graphical language and visualize the relation in a scatterplot

Models and Modeling in BI – Model Structures

- Putting all these things together leads to the concept of a ***model structure*** composed of:
  - Model Language:
    - Syntax defines basic elements and the rules how to compose model elements
    - Semantic defines the meaning of the elements in the language, independent from any domain
    - Notation for communication of the expressions in the language

Models and Modeling in BI – Model Structures

- Model Elements: Certain expressions in the model language, useful for describing facts about the business process
- Generic questions: Questions formulated in the semantic of the model language about properties of model elements
  - Generic questions can be answered by specific analysis techniques

Models and Modeling in BI – Modeling

- A mapping of some part of the domain semantic of a business process into a certain model structure ("***Conceptual Modeling***")
  - Examples for domain concepts and relations:
    - Health Care Use Case:
    - Higher Education Use Case:
    - CRM Use Case:

Models and Modeling in BI – Modeling

- Definition of a model configuration: admissible expression in a model structure which allows formulation of the analytical goal in questions about the model configuration
- Connection of model configuration with observations: data about the instances of the business process have to fit to the model configuration, i.e., views and perspectives
- Definition of model variability: Usually data are blurred due to noise or statistical variability

Models and Modeling in BI – Model Assessment and Quality

- Quality criteria for business process models
  - Correctness: model is syntactical correct and mapping of domain semantic and model semantic is appropriate
  - Relevance: model complies with intended function, i.e., explain past observations and predict future observations
  - Economic efficiency: trade-off between complexity and costs (Occams razor)
  - Clarity: model can be understood by users
  - Comparability: model fits in the overall analysis framework of the business process

Models and Modeling in BI – Model Assessment and Quality

- Quality criteria for empirical models
  - Objectivity: Results are independent of the person using the model
  - Reliability: results of the model can be reproduced
  - Validity: model is useful from a practical point of view
    - Content validity: model represents phenomenon under consideration
    - Criterion validity: high correlation between model results and other external properties
    - Construct validity: new results can be derived from model

Models and Modeling in BI – Models and Patterns

- Patterns describe local behavior whereas models describe global behavior
  - Examples:
    - Medical treatment process: a pattern of co-occurrence of certain medications
    - Customer relationship: A pattern of occurrence of certain combination of variables like outliers

Model Structures – Logical and Algebraic Structures

- Language:
  - Propositional logic and predicate logic
    - Individual constants (names), e.g., “`John Dee`” “`Business Intelligence`”
    - Variables: placeholders for constants, e.g., “`Student`”, “`Course`”
    - Functions: operating on constants or variables, e.g. “`grade(Student) = passed`”
    - Predicates: define properties for the individual constants, e.g., “`AttendsBI`”
    - Quantifiers (“`for all` (∀) ”, “`exists` (∃)”)

Model Structures – Logical and Algebraic Structures

- Language:
  - Propositional logic and predicate logic
    - Definition of terms by individual constants, individual variables, and functions
    - Generate atomic formulas by a predicate symbol followed by a number of terms for which the predicate is applicable, e.g. "`AttendsBI[John Dee]`"
    - Build well formed formulas using propositional calculus and quantifiers, e.g.,

```
∃(Student)(∀(Course) grade(Student,
Course) =passed)
```

Model Structures – Logical and Algebraic Structures

- Model elements and generic questions
  - Building expressions according to predicate logic
  - Assign truth values to the expressions (interpretation)
  - If the interpretation results in truth values TRUE for all possible assignments of the free variables we call the interpretation a model
  - Generic questions are whether a well formed formula is true

Model Structures – Logical and Algebraic Structures

- Modeling using logical structures tries to capture domain knowledge in a logical form
- The simplest form are terminology systems like taxonomies

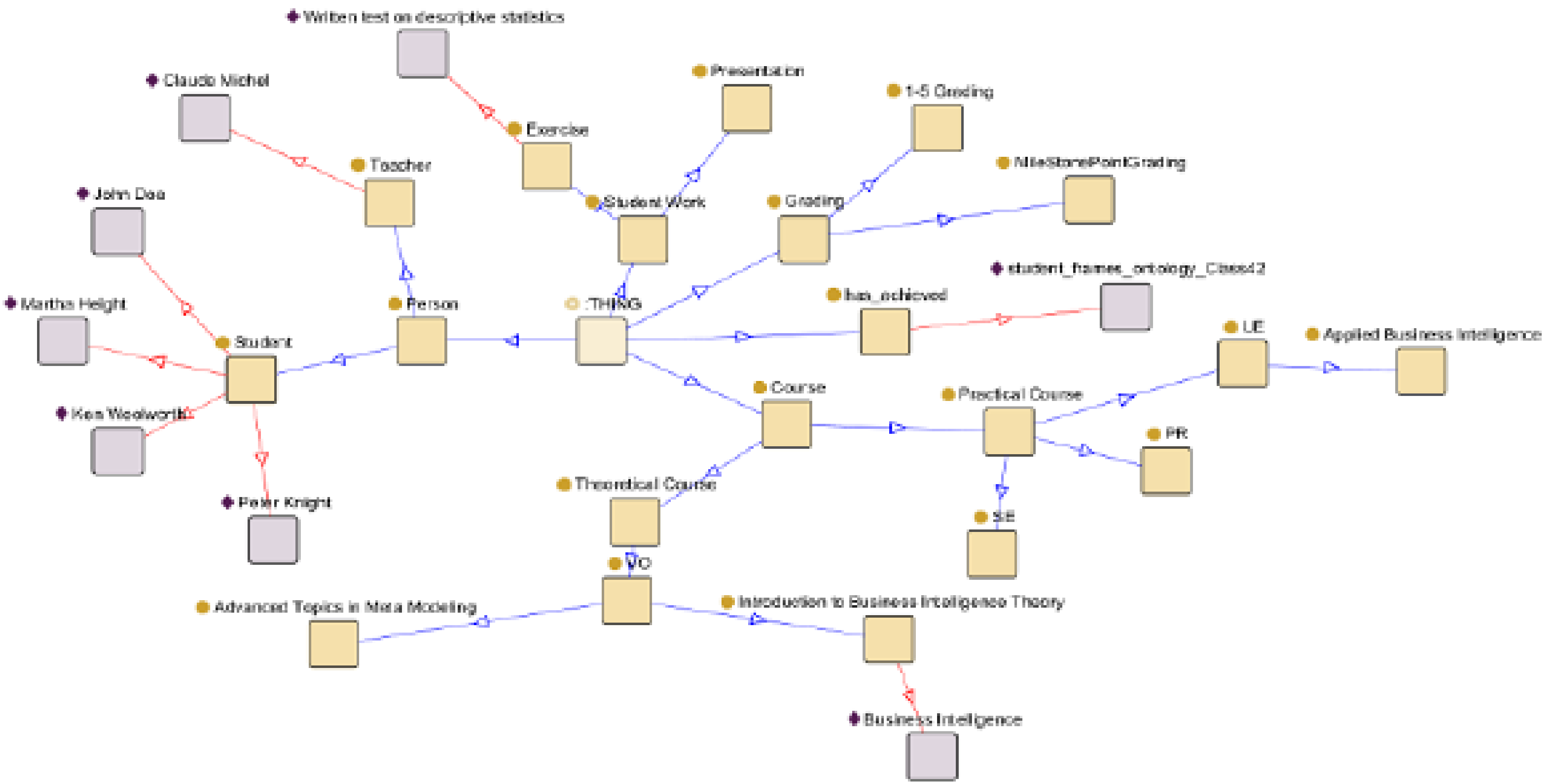
Model Structures – Logical and Algebraic Structures

- Ontologies:

  “A specification of a conceptualization”

  – OWL:
    - T-Box: Vocabulary of a domain as a logical theory
    - A-Box: Assertion about the domain, which has to be checked
    - Uses the open world assumption, i.e., anything can be entered in the T-Box unless it violates constraints

Model Structures –
Logical and Algebraic Structures



Model Structures – Logical and Algebraic Structures

- Frames
  - Representation in an object oriented style
    - For each object a number of slots are defined for attributes of the objects
  - Frames use the closed world assumption, i.e., a statement is true if its negation cannot be proven within the system
    - Example : “All birds can fly” (closed world)
      “There exist non flying birds” (open world)

Model Structures – Graphs

- Language:
  - Syntactic elements:
    - nodes (vertices)
    - Edges (directed, undirected)
    - Labels for edges (e.g., “distance”) or nodes (e.g., “degree”)
  - Notation:
    - Numeric representation (adjacency matrix)
    - Visual representation

Model Structures – Graphs

- Model elements:
  - Special kinds of graphs, e.g., trees, series parallel networks, bipartite graphs
  - Connected graphs (path)
- Generic questions
  - Generic questions refer to properties of the graph and can be answered by well known algorithms like spanning tree, shortest path, best matching of nodes

Model Structures – Graphs

- Modeling using graph structures
  - Business process modeling and notation (BPMN)
  - Petri nets

Model Structures – Analytical Models, Calculus

- Language:
  - Variables in one or more dimensions
  - Mathematical functions
- Model elements:
  - Classical functions (linear functions, logarithm, exponential functions,…)
  - Norm of a vector, distance

$$\| \mathbf{x} \| = \sum_{i=1}^{p} x_i^2 \, , \, d(\mathbf{x,z}) = \| \mathbf{x} - \mathbf{z} \| = \sum_{i=1}^{p} (x_i - z_i)^2$$

Model Structures – Analytical Models, Calculus

- Model elements:
  - Inner product

$$(\mathbf{x}^t \cdot \mathbf{z}) = \sum_{i=1}^{p} (x_i \cdot z_i)^2$$

  - Linear functions in more than one variable (matrices)

$$f(\mathbf{x}) = B\mathbf{x}$$

Model Structures – Analytical Models, Calculus

- Generic questions:
  - Properties of functions
  - Minimization and maximization of a function
    - Value of the minimum: $z = \min f(\mathbf{x})$
    - Argument of the minimum:

$$\mathbf{x}_m = \arg\min f(\mathbf{x}) \quad (f(\mathbf{x}_m) = z)$$

Model Structures – Analytical Models, Calculus

- Generic questions:
  - Matrix factorization: If C is a symmetric positive definite matrix (covariance matrix) then we can represent this matrix in the form:

$$C = P \cdot D \cdot P^t$$

Here D is a Diagonal matrix and P is a matrix with orthogonal columns

This is frequently used for dimensionality reduction

Model Structures – Analytical Models, Probability

- Language:
  - Events, Calculus of events: E
  - Probability of events

$$P(E), \quad odds(E) = \frac{P(E)}{1 - P(E)}$$

  - Random variables as model for measurement: X
  - Probability Distribution:
    - Distribution function: $F(x) = P(X \le x)$
    - Density function and probability function: $p(x)$

      We interpret the density as likelihood of an observation

Model Structures – Analytical Models, Probability

- Language:
  - Conditional probability and independence:

$$p(x \mid y) = p(x, y) / p(y)$$

  - Two variables are independent if

$$p(x, y) = p(x)\,p(y)$$

  - Bayes Theorem:

$$p(x \mid y) = p(y \mid x) / p(y)$$

  - Interpretation of Bayes Theorem in the discrete case: Compute column percentages from row percentages

Model Structures – Analytical Models, Statistics

- Language:
  - Statistical units (observation units)
  - Population
  - Observable variable
  - Transfer the concepts of probability to observations, e.g., “distribution” to “sample distribution” (“empirical distribution”)

Model Structures – Analytical Models, Statistics

- Model elements and generic questions:
  - Descriptive methods
  - Inferential methods
    - Estimation
    - Testing
    - Confidence regions
- Modeling methods
  - Regression

Models and Data – Data Generation

- In BI we have usually ***secondary data***, i.e., data which have been collected for other purposes
  - Transactional data
  - Administrative data
  - Web data,…
- An important question for interpretation of results is defining the population which is represented by the data (e.g., tweets or evaluations on portals)
- Measurement of the data

Models and Data – Temporal Aspects

Elements of the Knowledge Based Temporal Abstraction Method

- *Time stamps* T_i are the basic primitives with a predefined granularity and a well defined zero.
- Time intervals $T = [T_{start}, T_{end}]$ are defined as pairs of time stamps for start and end. Time points are zero length intervals.
- An *interpretation context* ξ is a proposition that can change the interpretation of parameters within the scope of a time interval. Interpretation contexts can be nested.
- A *context interval* $< \xi, I >$ defines time intervals for which the interpretation context holds.
- An *event proposition* e represents the occurrence of an external volitional action or process and has to be distinguished from a measurable datum.
- An *event interval* $< e, I >$ represents the temporal duration of an event e.
- A *parameter schema* π is a measurable aspect of the state of the world (states of a process) with values in some domain $v \in V_\pi$. Parameter schemas may be of different type: primitive parameters (measurable data), abstract parameters (concepts), constant parameters (instant specific or instant independent).

Models and Data – Temporal Aspects

- A *parameter proposition* $< \pi, v, \xi >$ defines the values of parameters in a context.
- An *abstraction function* $\theta \in \Theta$ maps parameters into abstract parameters.
- A *parameter interval* $< \pi, v, \xi, I >$ denotes the value v of the parameter π in the context ξ during time interval I.
- An *abstraction* is a parameter or a parameter interval.
- An *abstraction goal* $\psi \in \Psi$ represent a specific intention or goal.
- An *abstraction goal interval* $< \psi, I >$ represents the idea that abstraction goal ψ holds in interval I.
- *Induction of context intervals* allows the induction of events, parameters, or abstraction goal propositions for some context interval.

- Source: KBTA implemented as the RÉSUMÉ System: (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.7866&rep=rep1&type=pdf)

Models and Data - Quality

Quality dimensions

- Relevance: Relevance measures in how far the data are useful in the intended context.
- Accuracy: Accuracy is the degree of conformity of a measure to a standard or a true value.
- Completeness: Completeness is a characteristic measuring the degree to which all required data is known. with respect to depth, breath and scope.
- Timeliness: Data coming early or at the right time, appropriate or adapted to the times or the occasion.
- Consistency: Consistency is expressed as the degree to which a set of data is equivalent in redundant or distributed databases.
- Coherence: Coherence refers to the adequacy of the data to be reliable combined in different ways and for various uses.
- Reliability: Reliability is a characteristic of an information infrastructure to store and retrieve information in an accessible, secure, maintainable, and fast manner.