UNIT II

Modelling in BI

* Modelling & Models in BI

* Defn:

* Model:

Model represents some part of business process & allow precise formulation of interesting questions

- We can realize the model by using representation function.

* 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 phenomena.

1) Idealized models:
- e.g. -

control flow of the business process, a treatment process, a course design

2) Analogical Models:

Overtake ideas from other sciences.
- e.g.

gravity models for relations betn person in dependance of distance.

3) Phenomenological models:

statistics, ~~e.g~~
- e.g:-

regression.

- Models of Data:

In it, the task is to learn the most appropriate model (Machine Learning, Data Mining).
- e.g.-

Churn Management.
- which variable influence the churn behavior of a customer.

e.g:-

age, sex, marital status, income)

- Models of Theories:
- Each application domain of BI has specific domain knowledge, usually defined by concepts & relation (logical relation) betn the concepts.
- concepts & logical relns define a formal system (ontology)
- Understanding this formal system as a theory data instances are models of this theory.
- Database models.

- Formulation of Models:
- Each language has its own semantic allowing defn of certain model elements an formulation of generic questions.

- Queries of database
- simulation occurance of two events in a business process.
- Graph models for social n/w.
- E.g.

Relations betn attributes.

a) Formulate a query in a data model & represent the result as a table.

b) Define a regression model & formulate the reln as an eqn

c) Use a graphical lang. & visualize reln in scatterplot.

* Model Structures:

- composed of:

1) Model language
2) Model elements
3) Generic questions

1) Model language

- syntax defines basic elements & the rules how to compose model elements.
- semantic defines the meaning of the elements in the language independent from any domain.

2) Model Elements:

Certain expressions in the model language useful for describing facts about business process

3) Generic Questions.

Can be answered by specific analysis techniques.

* Modelling -

A mapping of some part of the domain semantics of a business process into certain model structure (conceptual mapping)

- E.g. for concepts & relns:

1) HealthCare Use case
2) Higher Education Use case
3) CRM Use case.

* Model Configuration: -

* Defn:

Admissible expression in a model structure which allows formulation of analytical goal in questions about model configuration.

* Model Assessment & Quality of Models:

- Quality Criteria for business process models.

* Correctness.

Model syntactical correct & mapping of domain semantic & model Semantic is appropriate

* Relevance -

Init explains past observations & predict future observations.

* Economic Efficiency -

Trade-off betn complexity & costs.

* clarity
Model can be understood by users.

* compatibility
Model fits in the overall analysis framework of the business process.

* Model Assessment & Quality of Models.

* Quality criteria for business process models:

1) Correctness -
Model is syntactical correct & mapping of domain semantic & model semantic is appropriate.

2) Relevance
Model complies with intended fn
i.e.
explain past observations & predict future observations.

3) Economic Efficiency
Trade-off betn complexity & costs.

4) clarity -
Model should be understood by users.

5) Compatibility
Model fits in the overall analysis framework of business process.

* Quality criteria for empirical models.

- Objectivity
Results are independant of the person using the model.

- Reliability
Results of the model can be reproduced.

- Validity
Model is useful from a practical point of view.

a) Content Validity
model represents phenomenon under consideration.

b) Criteria Validity
high correlation betn model results & other external properties.

c) Construct Validity
New results can be derived from model.

*Modelling using Logical structures: Ontologies & Frames

*Language: Propositional logic & predicate logic

- Individual const. (names) e.g. "John Dee"
- variables: placeholders for constants.
  e.g. "students", "course".
- Functions: Operating on constants or var.
  e.g. "grade(student = passed"
- Predicates: Define properties for individual const.
  e.g. "Attends BI"
- Quantifiers ("for all (∀)", "exists(∃)")
- Building expressions according to predicate logic
- Assign truth values to the expression
- If the interpretation results in truth values TRUE for all possible assignments of the free var. we call the interpretation model.
- Generic questions are whether a well formed formula is true.
- Modelling using logical structures tries to capture domain knowledge in a logical form.

*Ontologies:
A specification of 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.

*Frames:

- Representation in an object-oriented style
- For each object a no. of slots are defined for attributes of the object.
- Frames use the closed world assumption.
  i.e.
  a statement is true if its negation can't be proven within the system.
- E.g.
- All birds can fly". (closed world).
- There exist nonflying birds". (open world)

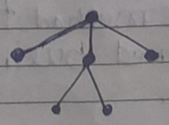
*Modelling Using Graph Structures: BPMN & Petri Nets.

- syntactic elements:
- Nodes (vertices)
- Edges (directed, undirected)
- Labels for edges (e.g. "distance") or nodes (e.g. days)

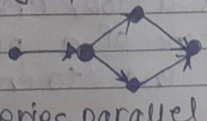
- Notations:
- Numeric representation (adjacency matrix).
- visual representation.

*Model Structure:

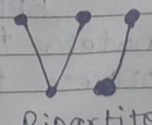
- special kinds of graphs e.g. trees, Series parallel n/w, bipartite graphs.
- connected graphs (path)



Tree



series-parallel graph(directed)

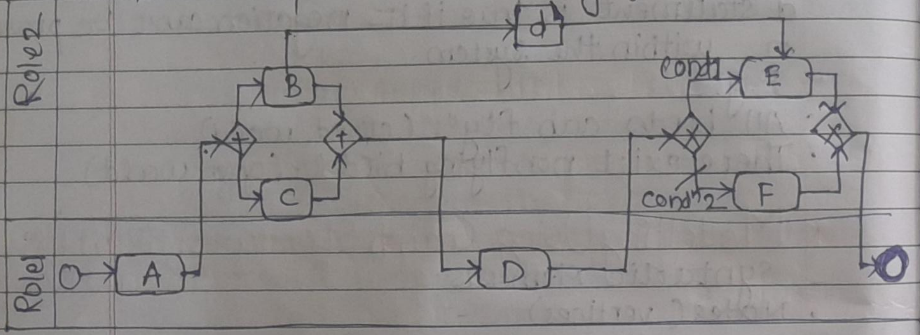


Bipartite graph.

- Generic Questions:

- Generic questions refer to properties of the graph & can be answered by well known algo. like spanning tree, shortest path, best matching of nodes.

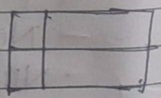
- Modelling using Graph structures:
- e.g.
- Business process modelling & notation (BPMN)



Control Flow — Data Flow — Events — Gateways — Organization

□ Task — ▯ Data Object — ○ Start — ◈ Parallel

→ Sequence Flow — → Data Association — ● End — ◈ Exclusive



cond Transition condn

* Defn

BPMN is a flow chart method that models the steps of a planned business process from end to end.

- It visually depicts a detailed sequence of business activities & info. flows needed to complete a process.

* BPMN elements types for business process diagram:

1) Flow objects: events, activities, gateways.
2) Connecting objects: sequence flow, message flow, association.
3) Swimlanes : Pool or lane
4) Artifacts : data object, group, annotation

1) Activity-
A particular task performed by person or system. It's shown by a rectangle with rounded corners.

2) Gateway-
Decision point that can adjust path based on condns or events. They are shown as diamonds. They can be exclusive or inclusive, parallel, complex.

3) Sequence flow-
Shows the order of activities to be performed. It's shown as a straight line with an arrow. It might show a conditional flow.

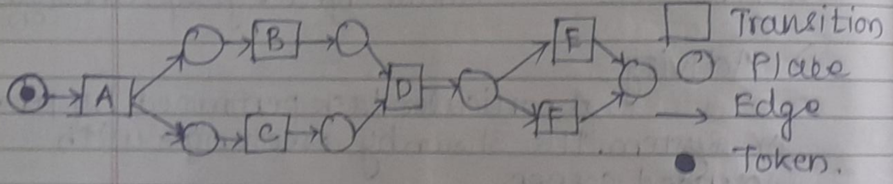
4) Message Flow:
Depicts messages that flow across "pools".

5) Association – Shown with a dotted lines, it associates an artifact or text to an event.

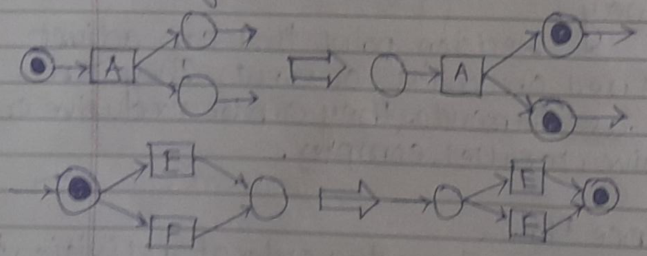
6) Pool & swimlane – It represents major participants in a process. It may be different company or department.

* Petri Nets.

* Process Model Represented as Petri net:



* Firing of Transitions:



* Modelling using Probabilistic Structure.

- Events, Calculus of events : E
- Probability of events P(E), odds(E) = $\frac{P(E)}{1-P(E)}$
- Random var. as model for measurement : X
- Probability Distribution.
  a) Distribution function : $F(x) = P(X < x)$
  b) Density function & f(x) probability f^n : p(x)
  c) We interpret the density as likelihood of an observation.
- Conditional Probability & independence
  $P(x|y) = p(x,y)/p(y)$
- Two var are independent if
  $p(x,y) = p(x) * p(y)$
  Baye's Them : $p(x|y) = p(y/x)/p(y)$
- Example:
  Joint Probabilities:

Usage Pattern	AgeGroup		Marginal
	young	Old	
high	0.2	0.1	0.3
moderate	0.3	0.2	0.5
inactive	0.1	0.1	0.2
marginal	0.6	0.4	1.0

* Modelling Using Analytical Structure:

* Calculus:
- var. in one or more dimensions.
- Mathematical f^ns: $f : X \to Y$, $y = f(x) = f(x_1, \ldots x_p)$

- Model Elements:
- Classical functions (linear f^ns, logarithms, exponential f^ns)
- Norm of a vector: $||x|| = \sqrt{\sum_{i=1}^{p} x_i^2}$
- Distance bet^n two vectors x & z:

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

- Inner Product, given a vector w of coefficients

$$f(x) = w^T x = w_1 x_1 + \cdots + w_p x_p$$

- Linear f^n in more than one var. (matrices)
  $f(x) = Bx$ where B is a $k \times p$ matrix.
- Projections

The orthogonal projections of a vector 'x' onto another vector w is defined by $P_w(x) = x' * \frac{w}{||w||}$

* Properties of functions:

- Minimization & maximization of a f^n
  - value of min: $z = \min f(x)$
  - Argument of minimum: $x_m = \arg\min f(x)$ $(f(x_m)) = z$
- Matrix factorization:

If 'C' is a symmetric positive define matrix (covariance matrix) then we can represent this matrix in the form

$$C = P * D * P^t$$

- Here 'D' is a diagonal matrix & P is a matrix with orthogonal columns.
- This is frequently used for dimensionality reduction.

* Models and Data:

* Data Generation:

- In BI we have usually secondary data i.e. data which have been collected for other purposes.

e.g.
- 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.

* Elements of the knowledge based temporal abstraction method:
- Time stamps T_i are the basic primitive with a predefined granularity & well defined zero
- Time intervals $T = [T_{start}, T_{end}]$ are defined as pairs of time stamps for start & 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 $\langle \xi, I \rangle$ defines time intervals for which the interpretation context holds.
- An event proposition 'e' represents the occurance of an external volitional action or process & has to be distinguished from a measurable datum.
- An event interval $\langle e, I \rangle$ represents the temporal duration of an event e.
- A parameter schema π is a measurable aspect of state of the world (states of a process) with values in some domain $v \in V_\pi$. Parameter schemas may be of different types.
  primitive parameters (measurable data),
  abstract parameters (concepts),
  Constant parameters (instant specific)
- A parameter proposition $\langle \pi, v, \xi \rangle$ defines value of parameters in a context.
- An abstraction function $\theta \in \Theta$ maps parameters into abstract parameters.
- A parameter interval $\langle \pi, v, \xi, I \rangle$ 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 $\psi \in \Psi$ represents a specific intention or goal.
- An abstraction goal interval $\langle \psi, I \rangle$ represents the idea that abstraction goal ψ holds in interval I.

- Induction of context intervals allows the induction of events, parameters or abstraction goalpropositions for some context interval.

* Quality Dimensions for Data.

- Relevance measures in how far the data are useful in the intended context.
- Accuracy is the degree of conformity of a measure to a standard or a true value
- Completeness is a characteristic measuring the degree to which all required data is known, w.r.t. depth, breadth, scope
- Timelinees: Data coming early or at the right time, appropriate or adapted to the times or the occasion.
- Consistency is expressed as the degree to which a set of data is equivalent in redundant or distributed databases.
- Coherence refers to the adequacy of the data to be reliable combined in diff'n ways & for various uses.
- Reliability is a characteristic of an information infrastructure to store & retrieve information in an accessible, secure, maintainable, fast manner.