Automatization of construction of dimension hierarchies of multidimensional data models

Redreev P.G.

Omsk Branch Institute of Mathematics SB RAS

OLAP technology basics

- OLAP – On-Line Analytical Processing
- Multidimensional data model – cube:
  - dimensions – textual
  - measures – numerical
  - dimension hierarchies are necessary for realizing roll-up and drill-down operations

Relational data model

CITY	STATE	POPULATION
San Diego	Texas	4490
Miami	Oklahoma	13880
Pittsburg	Iowa	509

Example of relation

Attributes: city, state, population

(San Diego,Texas,4490) – tuple

Functional dependencies, multivalued dependencies

«Composition table» data model

- «Composition table» forms from source relational database
- Schema of data model representation

$$Sch(R^*) = \{X, \bigcup_{i=1}^{N} Dom(Y_i) \times \{Z_i\}\}$$

X, Y_i, Z_i – sets of attributes. X, Y_i – generalized coordinates, Z_i – values in cells of two-dimensional table

«Composition table» data model

	Semester					
	1		2			
	Lectures	**Practical**	**Lectures**	**Practical**	**Examination**	**Test**
Subject name	*Hours amount*	*Hours amount*	*Hours amount*	*Hours amount*	*Semester*	*Semester*
Foreign language		3		2	4	1,2,3
Physical training		2		2		1,2,3,4
Domestic History	2	2			1,2	
Philosophy	2	2			1,2,3	
Economics			2	2	2	

University Curriculum

Hierarchies of attributes in $\boldsymbol{X}$ and $\boldsymbol{Y_i}$ ($\boldsymbol{i}$=1, 2,…, $\boldsymbol{N}$) defines order of arrangement of attrbute values in headings of strings and columns

Hierarchy schema

- Definition 1. A hierarchy schema is a directed, acyclic and weakly connected graph $\boldsymbol{H}=(\boldsymbol{A}, \boldsymbol{E})$, where $\boldsymbol{A}$ is a set of attributes, $\boldsymbol{E}$ is a set of arcs.
- Definition 2. Let $\boldsymbol{C}$, $\boldsymbol{D}$ be attributes. $\boldsymbol{H}$ is a hierarchy schema. $\boldsymbol{C} \prec \boldsymbol{D}$ if there is a path in $\boldsymbol{H}$ from $\boldsymbol{C}$ to $\boldsymbol{D}$.

Partial order

- For functional dependency $\boldsymbol{C \rightarrow D}$, $\boldsymbol{C_k \in C}$, $\boldsymbol{D_l \in D \; \forall \; k, l \; C_k \prec D_l}$
- For multivalued dependency $\boldsymbol{C \twoheadrightarrow D(E)}$, $\boldsymbol{C_k \in C}$, $\boldsymbol{I_l \in D \cup E \;\; \forall \; k, l \; I_l \prec C_k}$
- Order given by user

Construction of hierarchy schema

L is a set of dimension levels: attributes.

Step 1. Add arcs ***AB*** in ***H***, where $A \prec B$:

1. given by user.
2. got from functional dependency.
3. got from multivalued dependency.

Step 2. While graph ***H*** contains cycles, remove one arc from each cycle.

Step 3. Add in ***H*** vertices from ***L*** that are absent in ***H***. Add in ***H*** arcs if ***H*** is a disconnected graph.

Example for «University Curriculum» database

- ***L***={*semester number*, *studies type*}
- Multivalued dependency: *subject code*, *semester number* ↠ *studies type* (*semester arrangement*)
- *studies type* ≺ *semester number*
- Hierarchy schema formed by algorithm: