EIROPAS SOCIĀLAIS FONDS
EIROPAS SAVIENĪBA
IEGULDĪJUMS TAVĀ NĀKOTNĒ

This work has been supported
by ESF project No.
2009/0216/1DP/1.1.1.2.0/09/APIA/VIAA/044

On Implicitly Discovered OLAP Schema-Specific Preferences in Reporting Tool

Natalija Kozmina and Darja Solodovnikova

Faculty of Computing, University of Latvia

10th International Conference on Perspectives in Business Informatics Research

Riga, Latvia, 6-8 October 2011

Outline

- Motivation
- OLAP Reporting Tool
  - Reporting Metadata
  - Preferential Profile Metamodel
    - OLAP Preference Metadata
  - Logical Level Metadata
- Methods for Generation of Recommendations
  - Hot-Start Method
  - Cold-Start Method
- Conclusions

OLAP Personalization

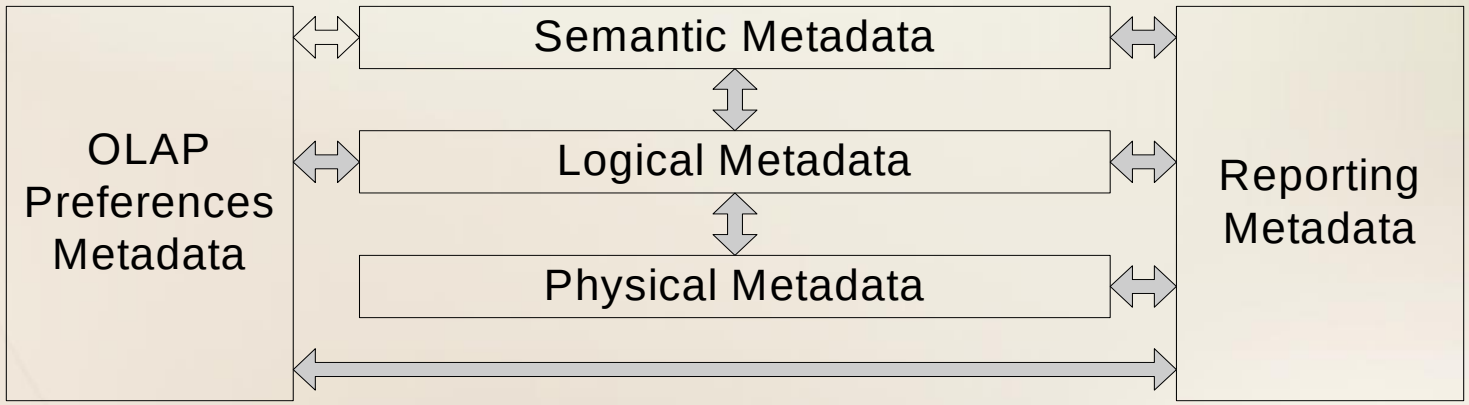
- Typical problems in DW field:
  - Large volumes of data,
  - Burdening data exploration,
  - Empty query result set,
  - While exploring previously unknown data, the OLAP query result may highly differ from expectations.
- Solution – introducing personalization in the field of data warehousing.

Motivation

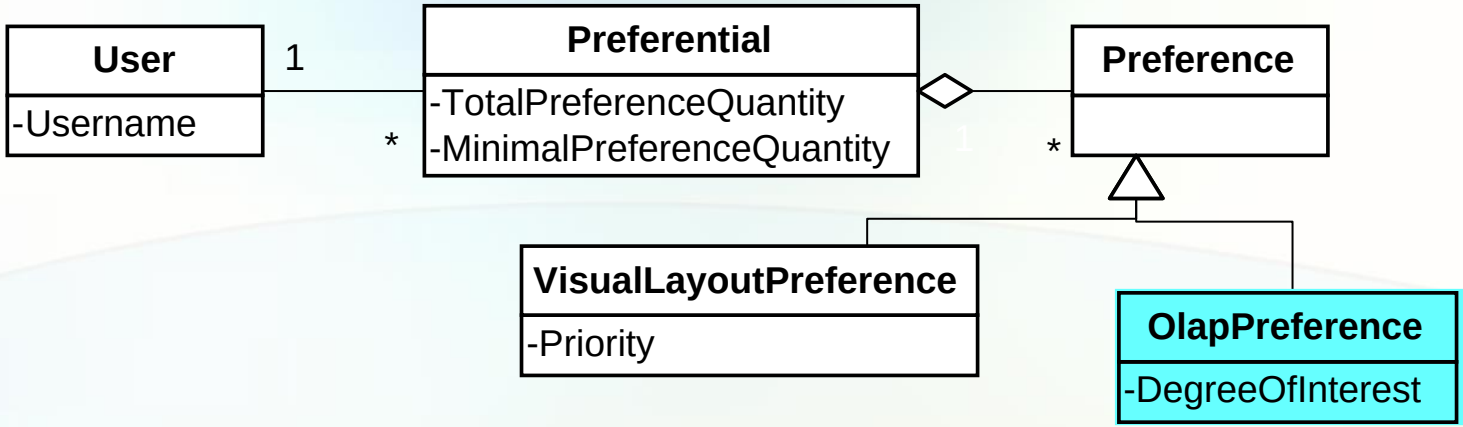
- OLAP reporting tool
- Different groups of users (e.g., students, professors, workers of the University, etc.)
- Each group or particular user has different…
  - rights, interests and skills,
  - reports' layout preferences.
- In this paper
  - We focus on acquiring user preferences implicitly to suggest a user reports that might be helpful.
  - We propose a way to orient in a variety of data warehouse reports, saving time and effort.

OLAP Reporting Tool

- Experimental environment: reporting tool developed at the University of Latvia.
- Operation of the OLAP reporting tool is based on metadata:
  - Logical: data warehouse schemata
  - Physical: storage of a data warehouse in relational database
  - Semantic: data stored in a data warehouse and data warehouse elements in a way that is understandable to users
  - Reporting: definitions of reports
  - OLAP preferences: definitions of user preferences on reports' structure and data.

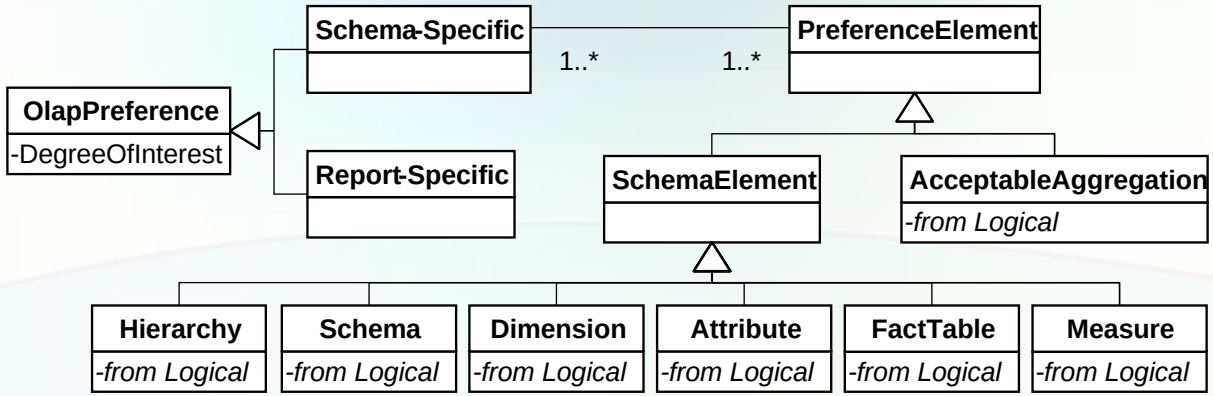


Preferential Profile Metamodel



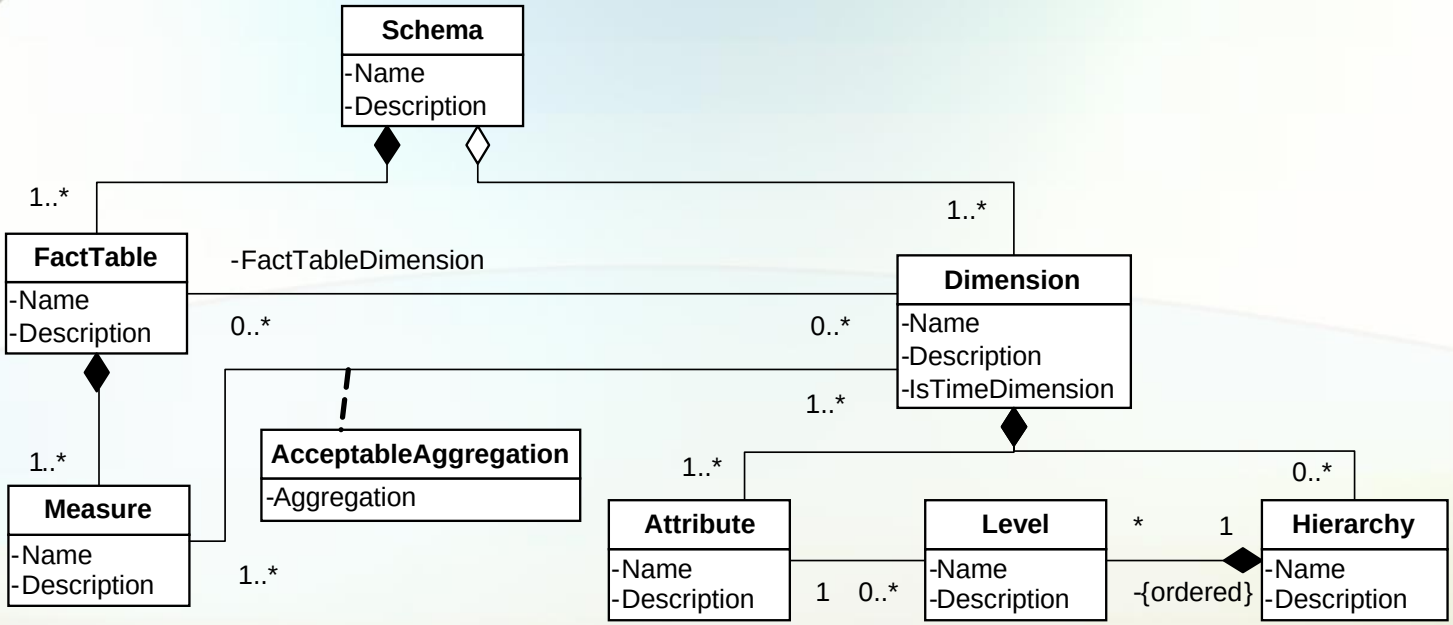
- *What* is the user expecting to get as a result?
- User preference modeling scenarios have been divided into two groups:
  - preferences for the contents and structure of reports (OLAP preferences),
  - visual layout preferences.
- Two ways of collecting user preferences:
  - explicitly (i.e., manually entered by user)
  - implicitly (i.e., analyzing user's activity by means of web-logs, visited links, etc.).

OLAP Preference Metadata



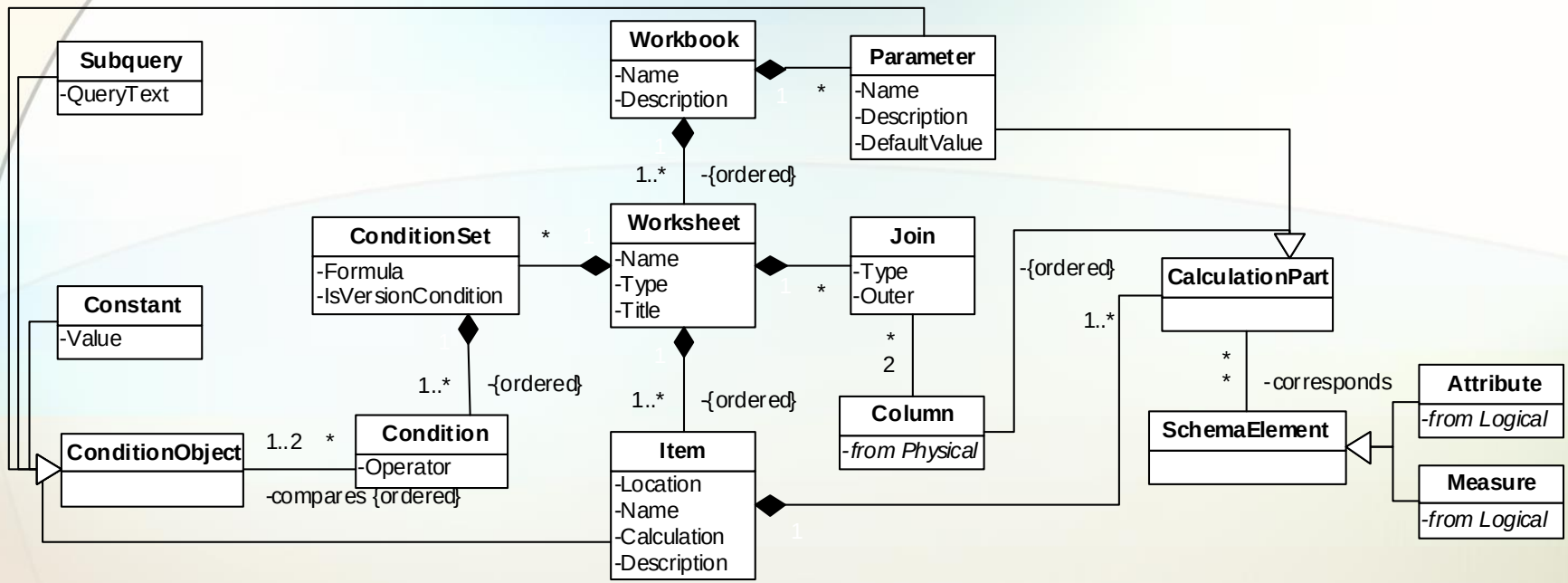
- Data about user preferences
- Preference contains user's degree of interest
- OLAP preferences:
  - Report-Specific preferences refer to preferences for particular reports and data restrictions in reports.
  - Schema-specific preferences are set for preference elements.

Logical Level Metadata



- Metadata at the logical level describes the multidimensional data warehouse schema.
- Data warehouse schema elements are included into the hierarchical structure:
  - A data warehouse schema is composed of interconnected fact tables and dimensions, which are composed of measures and attributes respectively.
  - Dimensions include hierarchies composed of ordered levels defined by attributes.
  - A fact table belongs to exactly one schema, but a dimension can be shared among multiple schemata.

Reporting Metadata



- Reporting metadata describes the structure of reports generated by users.
- Reports consist of
  - data items defined by computation formulas from parameters and table columns,
  - user-defined conditions and joins between tables.

Methods for Generation of Recommendations

- Hot-start method is applied for the user who has had a rich activity history with the reporting system.
- Cold-start method is applied, when
  a. a user of the reporting tool starts exploring the system for the first time,
  b. a user has previously logged in the system, but he/she has been rather passive.
- A borderline between the cold-start and the hot-start methods is defined by a *threshold*, which is the number of records in web-log appurtenant to a certain user.

Hot-Start Method

- *Step 1.* User preferences with degrees of interest (DOI) for data warehouse schema elements are discovered from the history of user's interaction with the reporting tool.
- *Step 2.* Reports that are composed of data warehouse schema elements, which are potentially the most interesting to a user, are determined.
- *Step 3.* Top-N potentially interesting reports are recommended to the user.

Hot-Start Method - Weight

- Weight of a schema $W(S_i)=2$.
- Weight of a fact table $W(F_i) = \frac{1}{n}$ (n is the number of fact tables belonging to one schema).
- Weight of a dimension in a schema equals to $W(D_i, S_j) = \frac{1}{k \cdot m_i}$ (n is the number of dimensions belonging to the schema S_j, $k = \sum_{l=1}^{n} \frac{1}{m_l}$, and m_i is the number of schemata, to which the dimension is related).
- Weight of a measure $W(M_i) = \frac{1}{n}$ (n is the number of measures belonging to the fact table).
- Weight of an attribute $W(A_i, D_j) = \frac{1}{n}$ (n is the number of attributes belonging to the dimension).
- Weight of an attribute, which is a level of a hierarchy $W(A_i, H_j) = \frac{W(A_i, D_k)}{n}$ (n is the number of attributes that make up levels of the hierarchy, and D_k is the dimension, to which the attribute belongs).
- **The weight of a schema element is equal to the sum of the weights of its subelements, except for hierarchies.**

Hot-Start Method

Discovering User Preferences - Algorithm

Input: User OLAP preferences for schema elements with the degrees of interest for each element and the schema element *E* used in a report. DOI(*SE*) is the user's degree of interest for the schema element *SE*, according to the user profile.

Output: User OLAP preferences with updated degrees of interest.

```
// if element E is a measure
if E instanceOf(Measure) then
   DOI(E)=DOI(E)+1;
   // getting a fact table, to which the measure E belongs
   F=getFactTable(E);
   DOI(F)=DOI(F)+W(E);
   // getting a schema, to which the fact table F belongs
   S=getSchema(F);
   DOI(S)=DOI(S)+W(F)*W(E);
// if element E is an attribute
else if E instanceOf(Attribute) then
   DOI(E)=DOI(E)+1;
   // getting a dimension, to which the attribute E belongs
   D=getDimension(E);
   // getting a schema, to which the dimension D belongs
   S=getSchema(D);
   DOI(D,S)=DOI(D,S)+W(E);
   DOI(S)=DOI(S)+W(D,S)*W(E);
   // getting hierarchies, levels of which correspond to the attribute E
   hierarchies=getHierarchies(E);
   foreach H in hierarchies do
      DOI(H)=DOI(H)+W(E,D)/countLevels(H);
   end
end
```

Hot-Start Method
Recommending Reports

- Content-based filtering approach is used.
- User's OLAP preferences are compared with schema elements used in each report to estimate the *hierarchical similarity* between a user profile and a report.

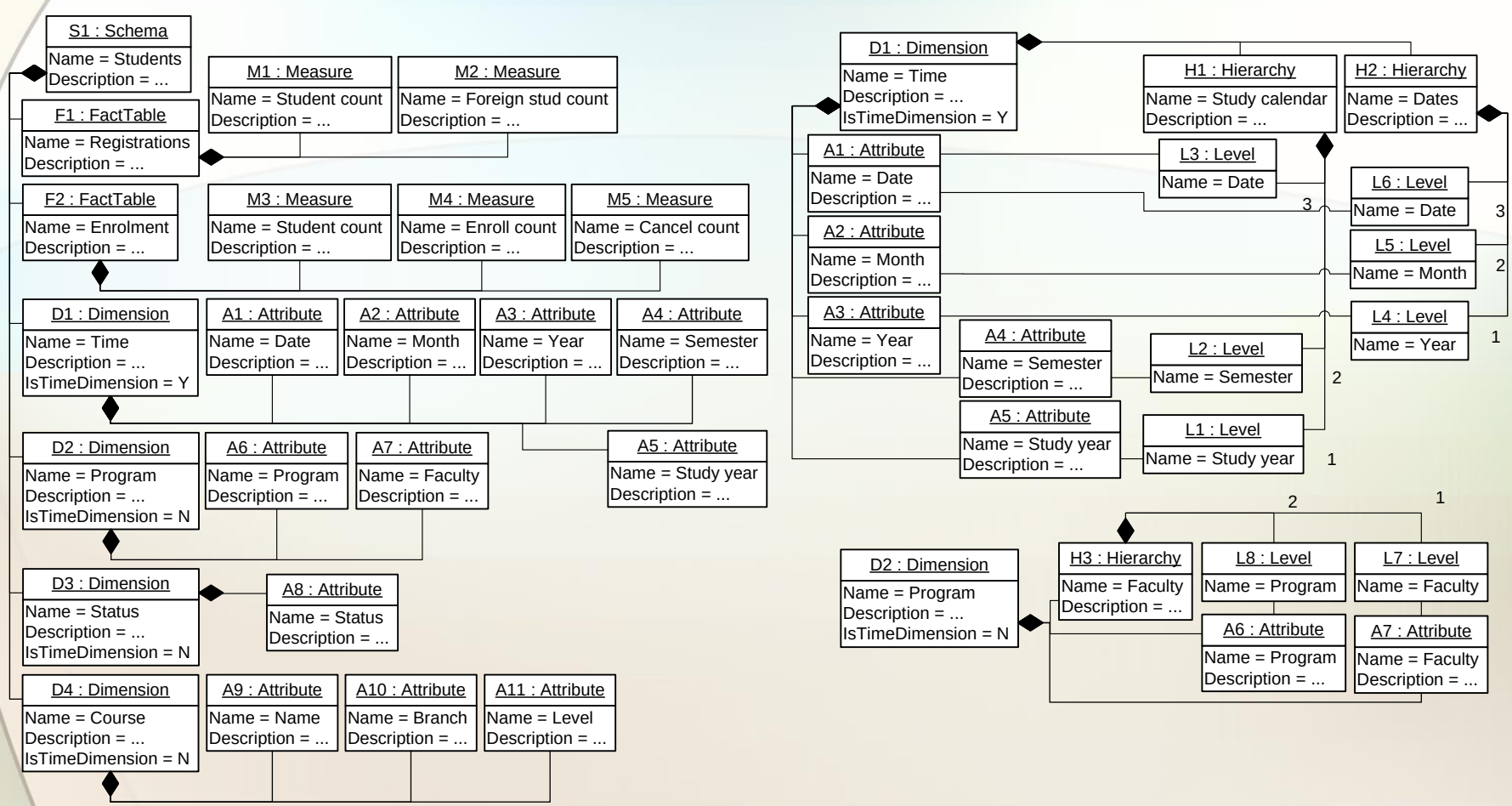
$$sim = \frac{\sum_{i=1}^{n} DOI(E_i)}{\sum_{j=1}^{m} DOI(G_j)}$$

where $E_1,\ldots,E_n$ are schema elements used in the report, and $G_1,\ldots,G_m$ are all schema elements in the user profile.

- Report recommendations:
  - In *fact-based* recommendations only those reports that contain measures from the fact tables with user's positive degree of interest are rated higher.
  - In *dimension-based* recommendations only those reports that contain attributes from the dimensions with user's positive degree of interest are rated higher.
- *Top-N* reports with the highest fact-based similarity and *Top-N* reports with the highest dimension-based similarity are recommended to the user.

Hot-Start Method

Example - Students data warehouse



Hot-Start Method

Example -Weights and Degree of Interest

	Schema	Fact tables		Measures					Dimensions				Attributes										
	S_1	F_1	F_2	M_1	M_2	M_3	M_4	M_5	D_1	D_2	D_3	D_4	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
Weight	2	1/2	1/2	1/2	1/2	1/3	1/3	1/3	1/13	4/13	4/13	4/13	1/5	1/5	1/5	1/5	1/5	1/2	1/2	1	1/3	1/3	1/3
DOI	4723/780	7/2	3	0	7	5	0	4	9/5	2	5	5/3	0	1	4	4	0	0	4	5	4	1	0

	Hierarchies			Attributes/Hierarchy Levels							
				Hierarchy H_1			Hierarchy H_2			Hierarchy H_3	
	H_1	H_2	H_3	A_5	A_4	A_1	A_3	A_2	A_1	A_7	A_6
Weight				1/15	1/15	1/15	1/15	1/15	1/15	1/4	1/4
DOI	4/15	1/3	1								

Hot-Start Method

Example -Weights and Degree of Interest

	Schema	Fact tables		Measures					Dimensions				Attributes										
	S_1	F_1	F_2	M_1	M_2	M_3	M_4	M_5	D_1	D_2	D_3	D_4	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
Weight	2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{13}$	$\frac{4}{13}$	$\frac{4}{13}$	$\frac{4}{13}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
DOI	$\frac{4723}{780}$	$\frac{7}{2}$	3	0	7	5	0	4	$\frac{9}{5}$	2	5	$\frac{5}{3}$	0	1	4	4	0	0	4	5	4	1	0

	Hierarchies			Attributes/Hierarchy Levels							
				Hierarchy H_1			Hierarchy H_2			Hierarchy H_3	
	H_1	H_2	H_3	A_5	A_4	A_1	A_3	A_2	A_1	A_7	A_6
Weight				$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{4}$	$\frac{1}{4}$
DOI	$\frac{4}{15}$	$\frac{1}{3}$	1								

Report R1:
Average foreign student count for each study program per semester

$$simD_{R1} = \frac{DOI(S_1) + DOI(D_2) + DOI(A_6) + DOI(H_3) + DOI(D_1) + DOI(A_4) + DOI(H_1)}{DOI(S_1) + DOI(F_1) + DOI(F_2) + DOI(M_1) + \ldots + DOI(H_3)} \approx 0.24$$

$$simF_{R1} = \frac{DOI(M_2) + DOI(F_1) + DOI(S_1)}{DOI(S_1) + DOI(F_1) + DOI(F_2) + DOI(M_1) + \ldots + DOI(H_3)} \approx 0.26$$

Hot-Start Method

Example -Weights and Degree of Interest

	Schema	Fact tables		Measures					Dimensions				Attributes										
	S_1	F_1	F_2	M_1	M_2	M_3	M_4	M_5	D_1	D_2	D_3	D_4	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
Weight	2	1/2	1/2	1/2	1/2	1/3	1/3	1/3	1/13	4/13	4/13	4/13	1/5	1/5	1/5	1/5	1/5	1/2	1/2	1	1/3	1/3	1/3
DOI	4723/780	7/2	3	0	7	5	0	4	9/5	2	5	5/3	0	1	4	4	0	0	4	5	4	1	0

	Hierarchies			Attributes/Hierarchy Levels							
				Hierarchy H_1			Hierarchy H_2			Hierarchy H_3	
	H_1	H_2	H_3	A_5	A_4	A_1	A_3	A_2	A_1	A_7	A_6
Weight				1/15	1/15	1/15	1/15	1/15	1/15	1/4	1/4
DOI	4/15	1/3	1								

Report R2:
Total student count enrolled into courses for each faculty per year

$$simD_{R2} = \frac{DOI(S_1) + DOI(D_2) + DOI(A_7) + DOI(H_3) + DOI(D_1) + DOI(A_3) + DOI(H_2)}{DOI(S_1) + DOI(F_1) + DOI(F_2) + DOI(M_1) + \ldots + DOI(H_3)} \approx 0.33$$

$$simF_{R2} = \frac{DOI(M_3) + DOI(F_2) + DOI(S_1)}{DOI(S_1) + DOI(F_1) + DOI(F_2) + DOI(M_1) + \ldots + DOI(H_3)} \approx 0.22$$

Cold-Start Method

- *Step 1.* Structural analysis of existing reports is performed.
- *Step 2.* Likeliness between two selected reports is revealed.
- *Step 3.* Top-N reports with the highest similarity values are shown to the user.

Cold-Start Method
Report Structure Vector

- Each report is represented as a *Report Structure Vector* (*RSV*):

$$RSV = ((e_{11}, e_{12}, \ldots, e_{1k_1}), \ldots, (e_{n1}, e_{n2}, \ldots, e_{nk_n}))$$

where e_{ik_i} is a vector coordinate, i.e., a binary value that indicates presence (equals 1) or absence (equals 0) of the instance of the report structure element, k_i is the number of elements in i-th structure, i is the index number of each structure ($i = 1, 2, \ldots, n$), n is the total number of distinct structure elements in reports.

	Attributes					Dimensions			Fact Tables		Measures			Acceptable Aggregation			Hierarchies			OLAP Schemas	
$\vec{r_1}$	1	1	1	0	…	1	1	…	1	…	1	0	…	1	0	…	1	1	…	1	…
$\vec{r_2}$	1	0	1	1	…	1	1	…	1	…	0	1	…	0	1	…	1	1	…	1	…
	year	semester	faculty	program		time	program		registrations		student	PhD student		AVG	COUNT		time	faculty		students	

$\vec{r_1}$ describes the structure of the report R1 – Average student count for each faculty per semester,

$\vec{r_2}$ describes the structure of the report R2 – Total PhD student count for each study program per year.

Cold-Start Method
Cosine / Vector Similarity

- Cosine/Vector similarity of the vectors $\vec{r}_1$ and $\vec{r}_2$

$$sim = \frac{\vec{r}_1 \cdot \vec{r}_2}{\left|\vec{r}_1\right| * \left|\vec{r}_2\right|}$$

where "." is the dot-product of two vectors and $\left|\vec{r}_i\right|$ is the length of each vector (i = 1, 2).

$$sim = \frac{\vec{r}_1 \cdot \vec{r}_2}{\left|\vec{r}_1\right| * \left|\vec{r}_2\right|} = \frac{8}{\sqrt{11} * \sqrt{11}} \approx 0{,}727$$

	Attributes					Dimensions			Fact Tables		Measures			Acceptable Aggregation			Hierarchies			OLAP Schemas	
$\vec{r}_1$	1	1	1	0	…	1	1	…	1	…	1	0	…	1	0	…	1	1	…	1	…
$\vec{r}_2$	1	0	1	1	…	1	1	…	1	…	0	1	…	0	1	…	1	1	…	1	…
	year	semester	faculty	program		time	program		registrations		student	PhD student		AVG	COUNT		time	faculty		students	

Cold-Start Method

- Discovering Similarities
  - The similarity is calculated among the active report (currently browsed by the user) and all the rest of the data warehouse reports.
  - *RSV* and *sim* values have to be recalculated dynamically when:
    - a new report is created
    - existing reports' structure is changed
- Recommending Reports
  - *Top-N* recommendations, i.e., links to the reports with *N* highest *sim* values sorted in descending order are shown to the user.

Conclusions

- Content-based methods for construction of recommendations for reports in the OLAP reporting tool:
  - Hot-start method defines user OLAP preferences in a reporting tool by means of analyzing user's past activity and determines reports that are composed of data warehouse schema elements, which are potentially the most interesting to a user.
  - Cold-start method examines the structure of the report being browsed at the moment and calculates the similarity of it with the rest of the reports.
- Future work
  - Estimation of the quality of recommendations for the group of users of reporting tools with different rights.
  - Extension of the approach by adding methods for
    - explicit definition and processing of OLAP and visual user preferences,
    - implicit handling of report-specific user preferences to state the most useful data in reports,
    - collecting and taking advantage of demographical information about users,
    - involving collaborative filtering.