POLITECNICO DI MILANO





Context-Aware Data Management

Emanuele Rabosio

Advisor: Elisa Quintarelli

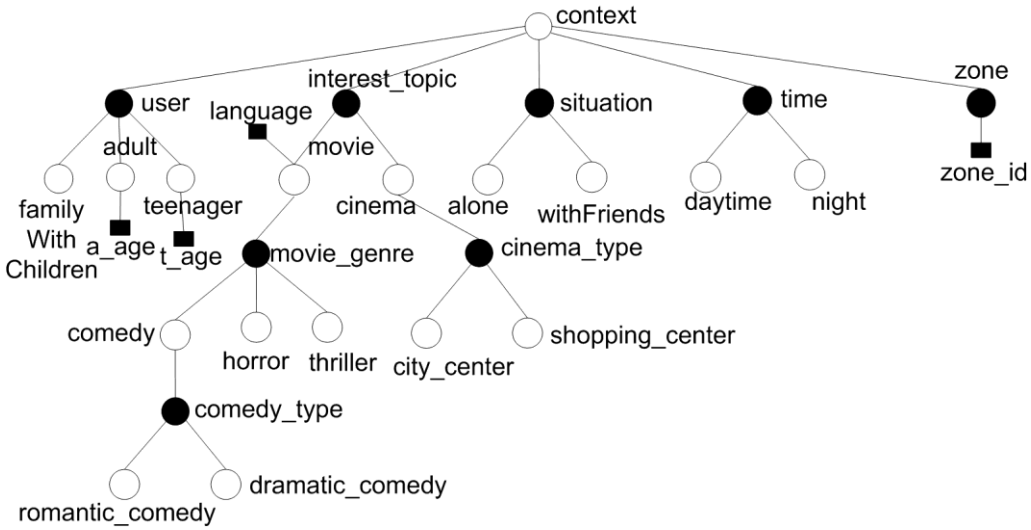
Dipartimento di Elettronica e Informazione

Introduction

- The current technological scenario is characterized by an extremely large variety of information sources, providing the users with an enormous amount of data
- This constitutes an unprecedented opportunity, but at the same time risks to confuse and overwhelm them
- A possible solution to this problem is **context-based data tailoring**: the system allows a user to access only the view that is relevant for his/her context
- In the literature several context models have been proposed, that usually describe the context through a series of dimensions

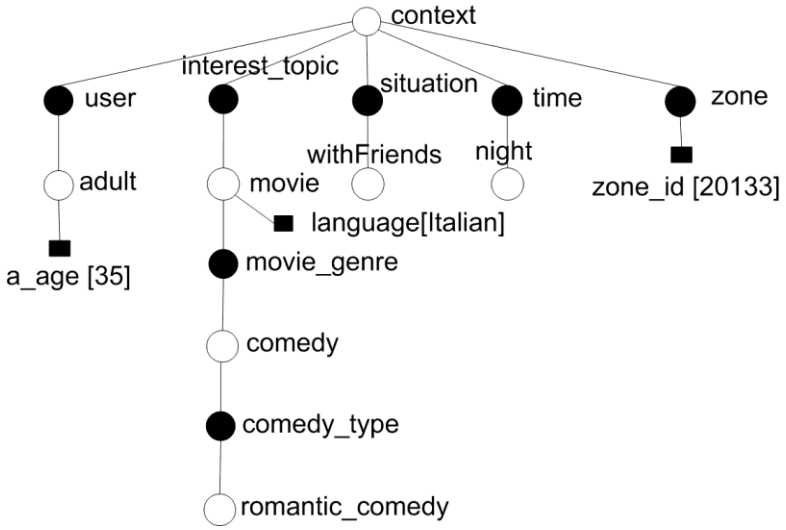
Context model: Context Dimension Tree (CDT)

- Provides a tree-based representation of context schemas and instances
- A context instance is a subtree of the related schema



Context schema

A context instance



Research issues

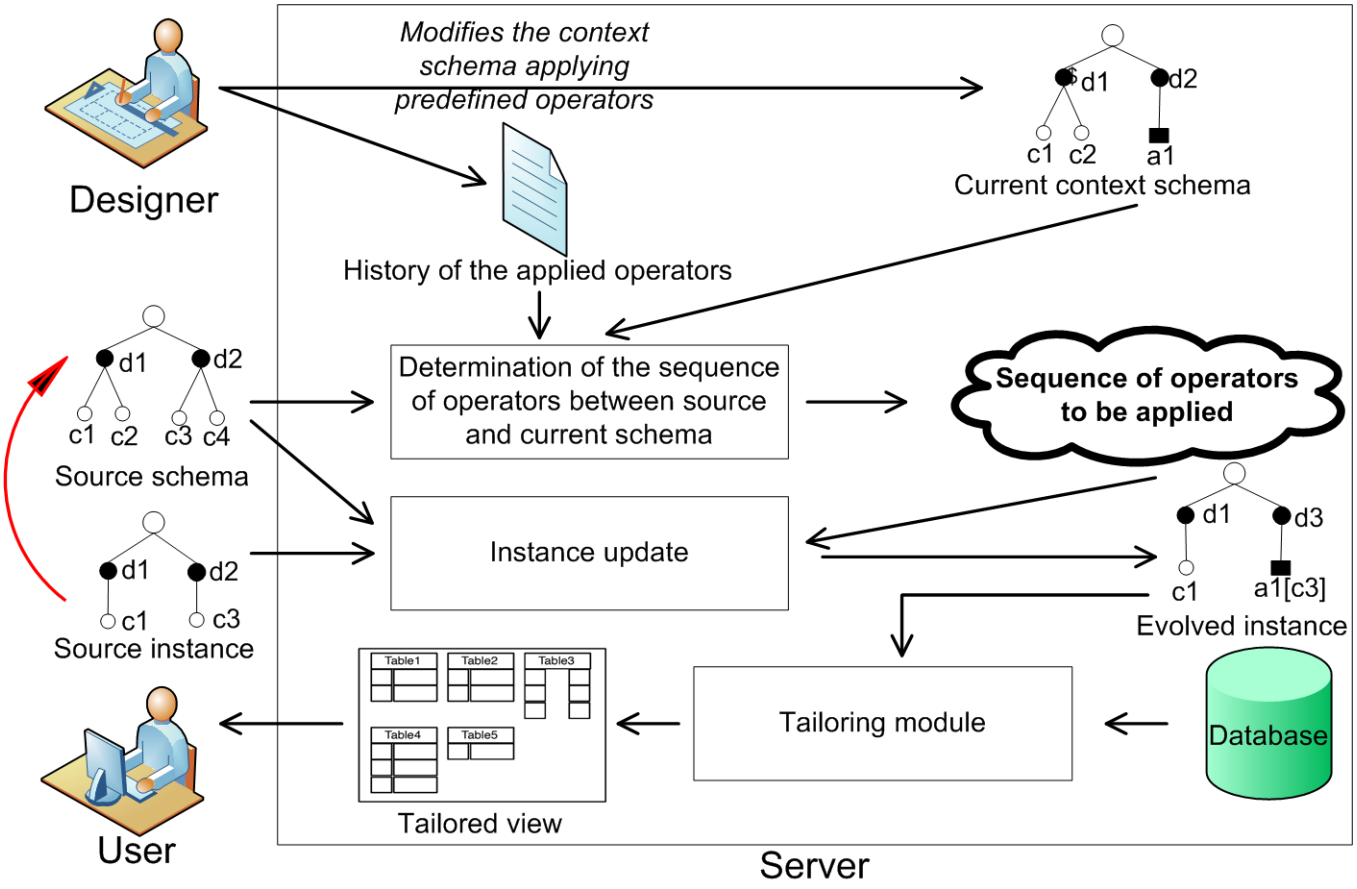
- In this PhD research we start by considering the CDT model with the aim of studying some context-related issues that we feel particularly relevant within data management:
  - **Context schema evolution**
  - Automatic **mining of context-aware preferences**

Context schema evolution

- The context dimensions and their values (together constituting the **context schema**) useful for data tailoring depend on the application requirements
- Application requirements are intrinsically dynamic and thus can change
- The change of requirements can be due to various reasons:
  - Changes in business policies
  - Market developments
  - Technology developments
- The changes in the application requirements lead to **context schema evolution**

→ *facilitate the modification of the context schema making the contexts defined according to old schemas still utilizable*

Framework for context schema evolution



Evolution operators

- An **update operation** *op* is implemented by two operators:
  - A **schema evolution operator** SU_{op}
  - An **instance evolution operator** IU_{op}
- The schema operators, employed by the designer to modify the schema, are characterized by a set of **preconditions** imposing restrictions on the source schema on which they are applicable
- The instance operators adapt the instances to the new schema, trying to preserve as much information as possible

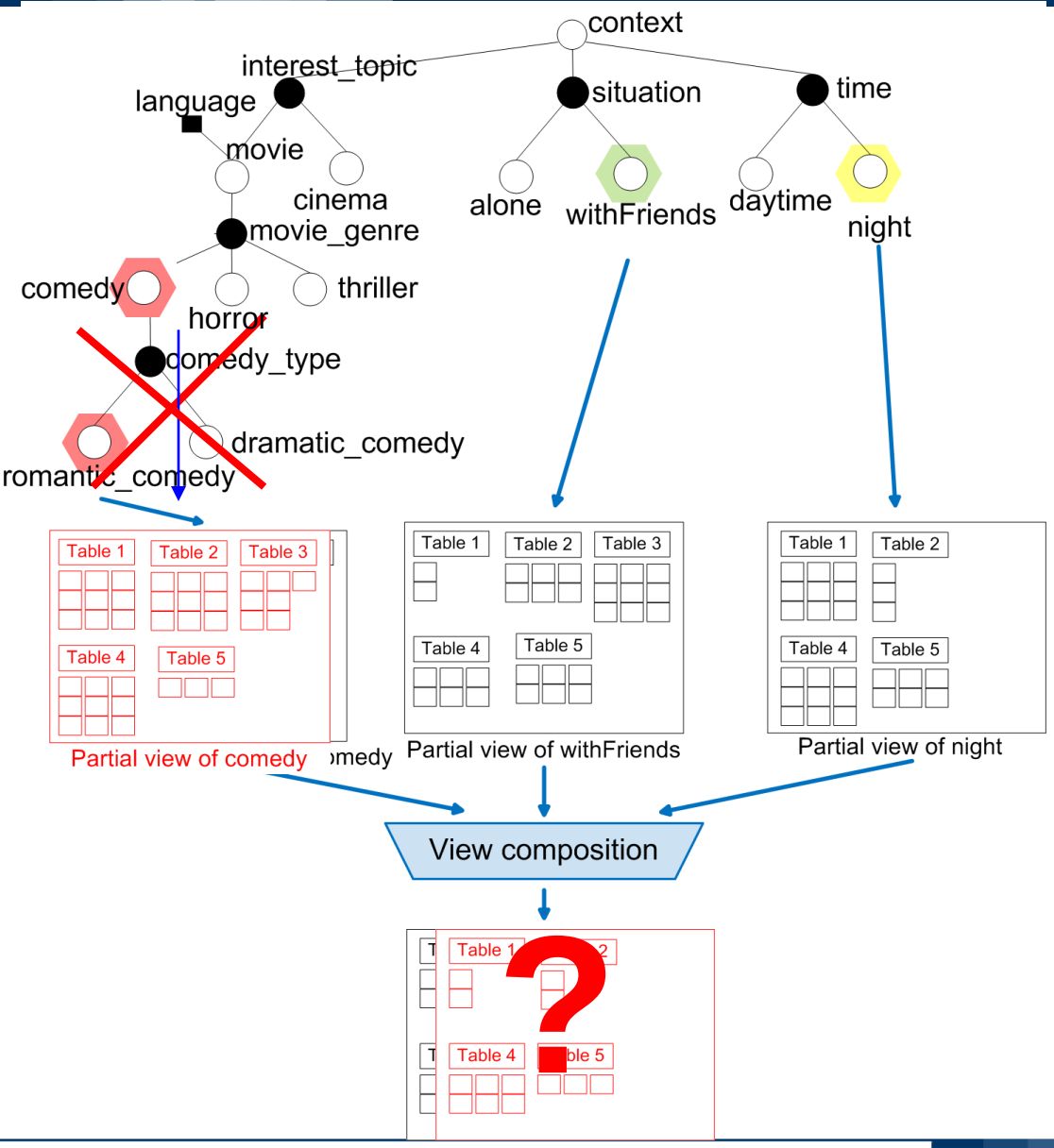
- Two categories of operators:
  - **Atomic operators**:
    - *Minimal*: each operator cannot be obtained as a combination of other ones
    - *Complete*: allow to evolve to any valid target context schema
  - **High-level operators**: compactly express common evolution needs

Evolution operators (2)

- We define **atomic operators** able to:
  - Delete subtrees
  - Insert subtrees
  - Replace sibling nodes with a new node
  - Replace subtrees with an attribute
  - Re-insert previously removed subtrees

- **High-level operators** allow to:
  - Move subtrees
  - Rename nodes or attributes
  - Insert attributes
  - Delete attributes

Node-based view definition



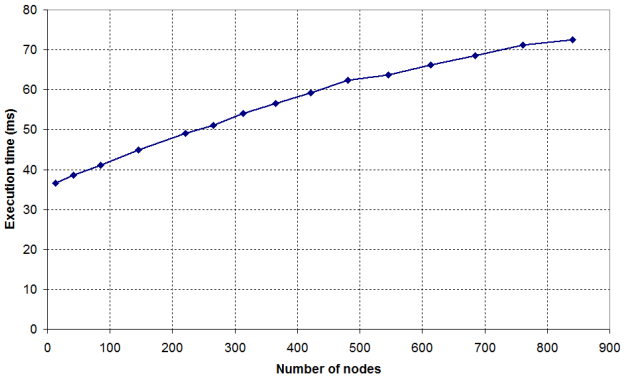
Requirement change:
no more useful to distinguish between romantic and dramatic comedies

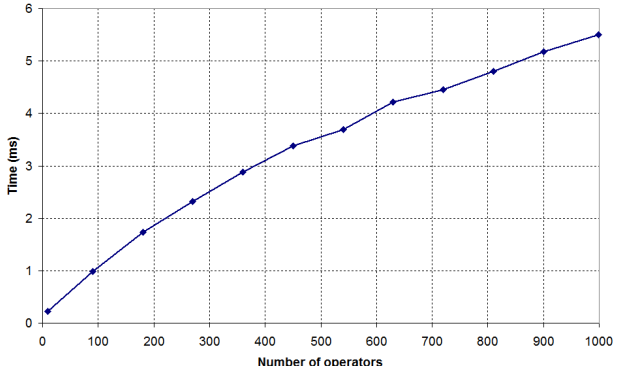
Delete (comedy_type)

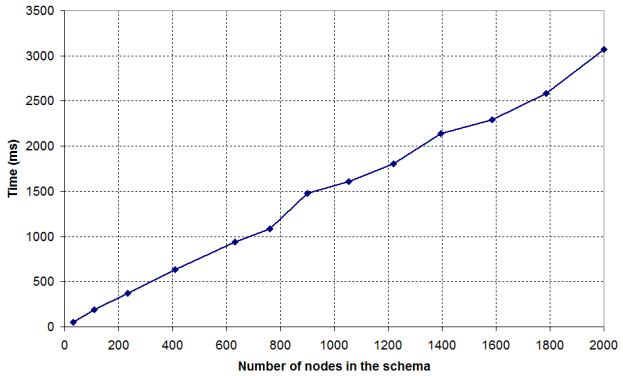
The designer defines a view for *comedy*

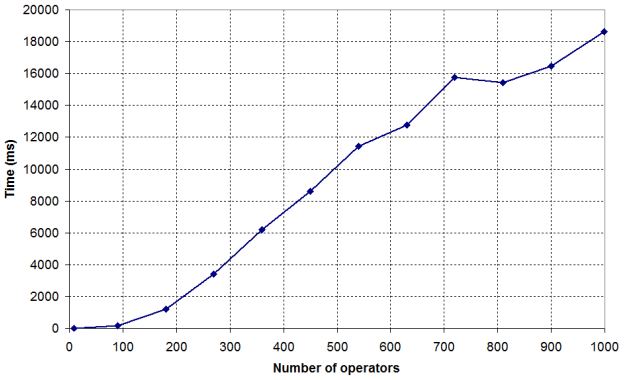
System implementation

- We have implemented the described techniques in a **Java prototype**, able to:
  - Check the preconditions and apply operators on schemas
  - Apply sequences of operators on instances
  - Optimize sequences of operators
- We have measured **execution times** to assess the feasibility of the system
- Also for extremely long sequences and huge schemas the times are **very low**









Automatic mining of context-aware preferences

- Context allows only a coarse-grained data tailoring
- The view associated with a context may be further refined exploiting the **user preferences** in that context
- Preferences are context-aware: a user might be very interested in comedies when he/she is alone, and in thrillers when he/she is with his/her friends
- The number of possible contexts may be huge also for a small CDT, therefore the user may be discouraged w.r.t. specifying preferences for all such contexts

→ *in this research we propose a methodology to automatically infer context-aware preferences*

Preference model

- Numerical scores on data tuples and attributes of a relational database, expressed in the range [0;1]

- **σ-preference** $P_\sigma(R)$ on the relation $R(X)$ is defined as <C, SQ, S>

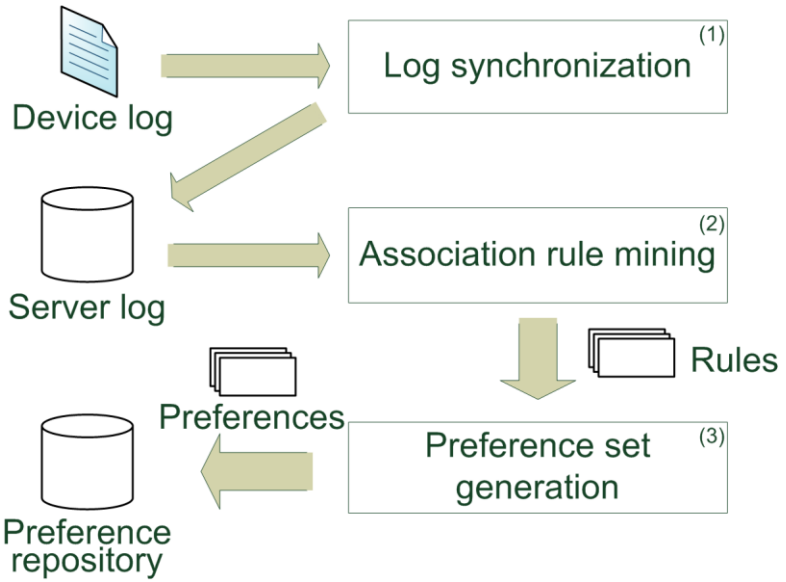
$$C = \texttt{interest\_topic} = \texttt{movie} \wedge \texttt{situation} = \texttt{alone}$$
$$SQ = \sigma_{\text{genre}=\text{“comedy”}}\,\text{MOVIE}$$
$$S = 0.8$$

- **π-preference** $P_\pi(R)$ on the relation $R(X)$ is defined as <C, A, S>

$$C = \texttt{interest\_topic} = \texttt{cinema} \wedge \texttt{situation} = \texttt{alone}$$
$$A = cinema.city$$
$$S = 0.9$$

Preference mining – Overview

- Objective:
  - Generate preference profiles analyzing log data, extracting knowledge in terms of association rules
- Input:
  - User activity log
- Output:
  - σ- and π-preferences
- The steps are performed independently for σ- and π-preferences
- Some minor tasks are omitted for brevity



σ-preference mining

Context: situation=alone, cinema_type=city_center
Query:

```
SELECT title
FROM movie
WHERE genre='comedy'
```

MOVIE			
movie_id	title	genre	year
m1	Movie1	comedy	2002
m2	Movie2	horror	2002
m3	Movie3	comedy	2007

Log of MOVIE:

Context dimensions — Movie attributes

id	situation	interest_topic	type	...	movie_id	title	genre	year
1	alone	cinema	city_center		m1	Movie1	comedy	2002
2	alone	cinema	city_center		m3	Movie3	comedy	2007

- This log is used to mine **association rules** connecting contexts with conditions on data

$$< situation = alone \wedge interest\text{-}topic = movie \rightarrow movie.genre = 'comedy', 0.7 >$$

- **Preference scores** are obtained combining the **confidence** of the rules and the **frequencies** of the related data in the dataset accessed by the user

π-preference mining

```
SELECT movie.title,
       movie.year
FROM movie
WHERE movie.genre='comedy'
```

```
SELECT actor.lastname
FROM actor
WHERE actor.citizenship='Italian'
```

Context dimensions

Attributes of the tables

situation	int_topic	type	m.movie_id	m.year	m.title	m.genre	a.actor_id	a.lastn	a.citiz
alone	cinema	city_center		1	1	1			
alone	cinema	city_center						1	1

- This log is used to mine **association rules**, connecting contexts with attributes

$$< situation = alone \wedge interest\text{-}topic = movie \rightarrow movie.genre, 0.7 >$$

- In this case the confidence of the rule is an indicator of preference, without requiring frequency computations

Experimental evaluation

- We have employed a **video-on-demand database**, containing information about movies, actors and directors
- Two **real log files** storing the activities of several users, associated with their context:
  - A commercial log of a video-on-demand company, with context specified through the dimensions day and time
  - A log collected through a web application, with richer context information (day, time and situation)
- The logs have been used to mine preferences, and compute the **recall**

- Results:
  - Our methods outperform other tested methodologies in terms of recall
  - Our methods outperform the same technique if run without considering the context information; the advantages are stronger when the context information is richer

Possible extensions

- Our work might be extended along several dimensions, applying techniques coming from other areas in the context-related field:
  - **Context sharing**, applying *data exchange* techniques
  - **Context-aware recommender systems**
  - Efficient formalization of the CDT through techniques of **formal verification**