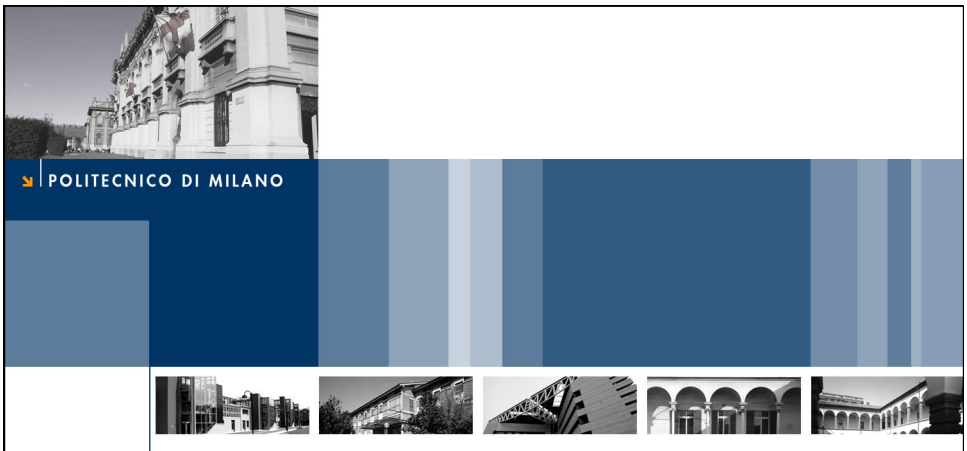


POLITECNICO DI MILANO

Managing Information Overload

Elisa Quintarelli

Elisa Quintarelli – 22th April 2015

The Challenges for Modern Information Systems

2

- Organizations grow and generate more information: they capture billions of bytes of information about their customers, suppliers and operations
- The pervasiveness of digital technologies has changed the way individuals interact with the external world (sensor technologies) and with one another (social media), generating a huge mass of content

→ Data has become as a torrent that flows through all possible digital channels.

POLITECNICO DI MILANO

Information Overload

3

- The term was used by Alvin Toffler in his book *Future Shock*, already back in 1970
- It refers to the difficulty of *understanding and making decisions* when too much information is available
- This is the main challenge presented by "Big Data"

Data Management: What does it mean today?

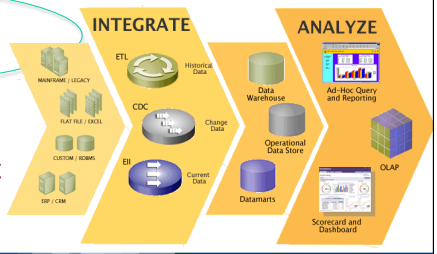
4

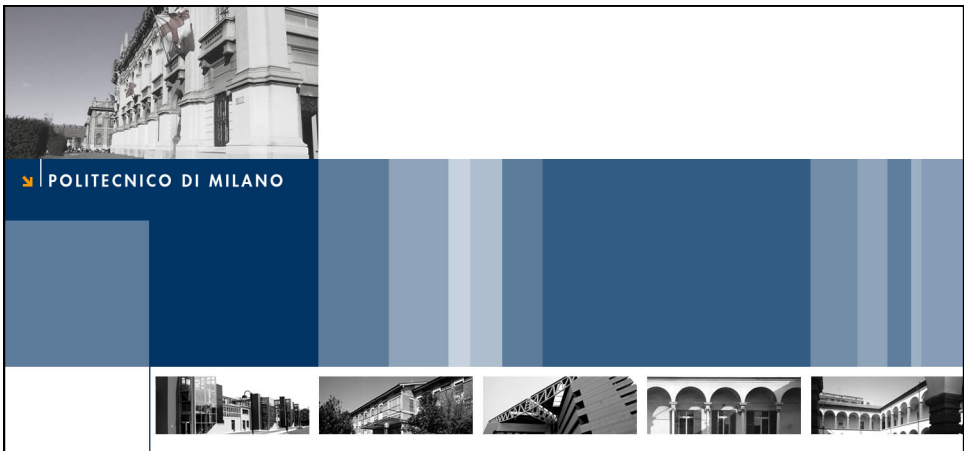
What do users want from us?

- Massive data integration/exchange
- Heterogeneity/mobility
- Incompleteness/uncertainty
- Interaction with the real-world
- Knowledge representation and reasoning
- **Massive data analysis and data mining**

- Summarized answers to user queries
- Personalization and context awareness

- Making sense of all this data: extract **useful knowledge**





Context-aware data personalization

Antonio Miele, Elisa Quintarelli, Emanuele Rabosio, Letizia Tanca

Elisa Quintarelli – 22th April 2015

Introduction

6

- Technological limitations of widespread small mobile devices (e.g.: pda, smart phones, …) constrain the amount of data which can be loaded on it
- There is a quest for reduction and personalization of the information each user accesses to
- Two main issues:
  - Reduce information noise
  - Satisfy device memory limitations

POLITECNICO DI MILANO

Background and Motivations

7

- The Context-ADDICT methodology tackles the presented problem, performing **context-based data tailoring**
- Current limitations:
  - The customization does not take into account **user preferences**
    - The context is the only driver
  - Data reduction to fit the memory of the device is not supported
    - No memory occupation model is considered
    - The approach is coarse-grain

The proposed approach

8

- Preferences express interests on data as numerical scores or explicit ordering relations
- Data scoring is commonly used to rank information in several of today's data management applications and search engines

- Fine-grained personalization of data can be performed by means of preferences

Contextual preferences – Examples

9

- A user may prefer spicy dishes at dinner, but not spicy ones at lunch
- A user may prefer comedy movies when he/she is alone, but thrillers when with his/her friends

Running Example

10

- An application for an integrated service of meal order and delivery is considered; it involves several restaurants and meal delivery taxi companies

CUISINES(cuisine_id, description)

DISHES(dish_id, description, isVegetarian, isSpicy, isMildSpicy, wasFrozen, category_id)

RESERVATIONS(reservation_id, customer_id, restaurant_id, date, time)

RESTAURANTS(restaurant_id, name, address, zipcode, city, state, zone_id, rnnumber, phone, fax, email, website, openinghourslunch, openinghoursdinner, closingday, capacity, parking, minimumorder, rating)

RESTAURANT_DISH(restaurant_id, dish_id)

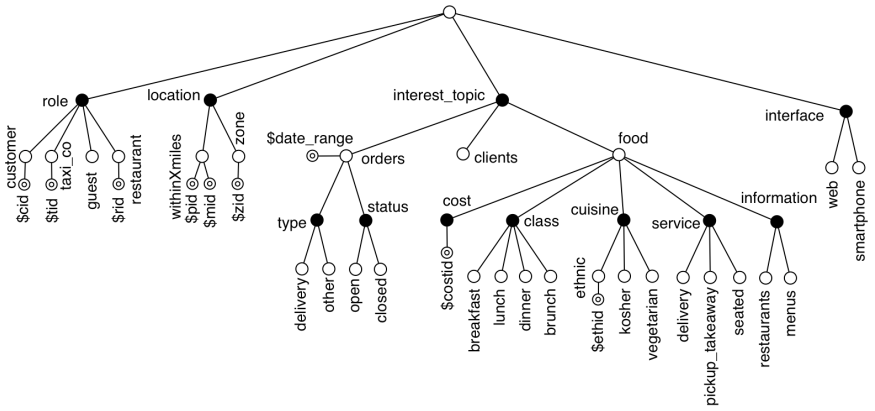
RESTAURANT_CUISINE(restaurant_id, cuisine_id)

RESTAURANT_SERVICE(restaurant_id, service_id)

SERVICES(service_id, name, description)

Context Dimension Tree of the running example

11



POLITECNICO DI MILANO

Preference Model – Overview

12

- Quantitative preferences
  - Numerical scores expressed in the range [0;1]
- Two different types of preferences:
  - σ-preferences
    - Expressed on data tuples
    - Acting in horizontal on each table
  - π-preferences
    - Expressed on relation attributes
    - Acting in vertical on each table

POLITECNICO DI MILANO

Preference Model – Preferences on data tuples

13

- **σ-preference** $P_\sigma(R)$ on the relation $R(X)$ is defined as

$$\langle C, SQ_\sigma, S_\sigma \rangle$$

- C : the context configuration
- SQ_σ : the selection query composed by a selection on $r(X)$ possibly semi-joined on selections of related tables (only on foreign keys)
- S_σ : the numerical score

Preference Model – Preference on data tuples Example

14

$$
\begin{array}{lllll}
CP_1 = & \langle C_1 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}), \\
 & SQ_{\sigma 1} & = & \sigma_{\texttt{isVegeterian}=1}(\texttt{dishes}), \\
 & S_{\sigma 1} & = & 0.1\rangle \\
CP_2 = & \langle C_2 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}), \\
 & SQ_{\sigma 2} & = & \sigma_{\texttt{isSpicy}=1}(\texttt{dishes}), \\
 & S_{\sigma 2} & = & 1\rangle \\
CP_3 = & \langle C_3 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}), \\
 & SQ_{\sigma 3} & = & \texttt{restaurant} \ltimes \texttt{restaurant\_cuisine} \ltimes \sigma_{\texttt{cuisine.description}=\text{“Chinese”}}\texttt{cuisine}, \\
 & S_{\sigma 3} & = & 1\rangle \\
CP_4 = & \langle C_4 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}), \\
 & SQ_{\sigma 4} & = & \texttt{restaurant} \ltimes \texttt{restaurant\_cuisine} \ltimes \sigma_{\texttt{cuisine.description}=\text{“Indian”}}\texttt{cuisine}, \\
 & S_{\sigma 4} & = & 0.3\rangle
\end{array}
$$

Preference Model – Preferences on relation attributes

15

- **π-preference** $P_\pi(R)$ on the relation $R(X)$ is defined as

$$\langle C, A_\pi, S_\pi \rangle$$

- C : the context configuration
- A_π : the list of attributes on the relation $R(X)$
- S_π : the numerical score

POLITECNICO DI MILANO

Preference Model – Preferences on relation attributes Example

16

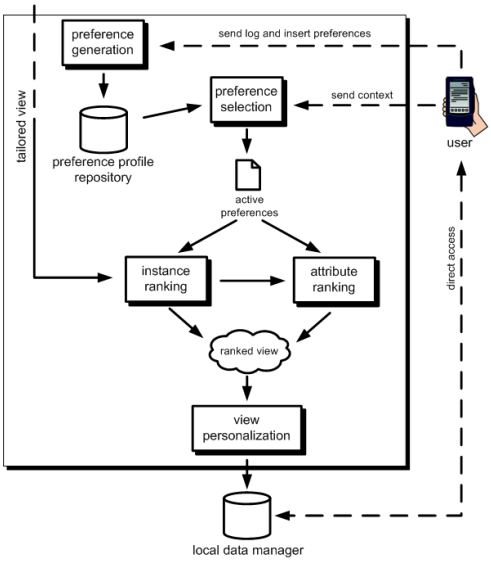
$$\begin{array}{lllll}
CP_5 = & \langle C_5 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \\
 & & & \wedge \texttt{interface} = \texttt{smartphone}, \\
 & A_{\pi 5} & = & \{\texttt{name}, \texttt{phone}, \texttt{zipcode}\}, \\
 & S_{\pi 5} & = & 1\rangle \\
CP_6 = & \langle C_6 & = & \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \\
 & & & \wedge \texttt{interface} = \texttt{smartphone}, \\
 & A_{\pi 6} & = & \{\texttt{address}, \texttt{city}, \texttt{rnnumber}, \texttt{fax}, \texttt{email}, \texttt{website}\}, \\
 & S_{\pi 6} & = & 0.2\rangle
\end{array}$$

POLITECNICO DI MILANO

Personalization Methodological Flow

17

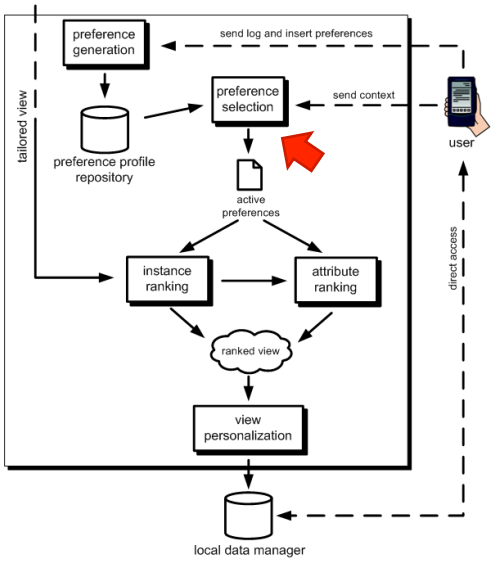
- Input:
  - The current context
  - The view tailored by the designer
  - The preference profile
- Output:
  - The personalized view
- Tasks:
  - Selection of active preferences
  - Ranking of the view attributes
  - Ranking of the view tuples
  - Personalization of the view



Preference Selection – Overview

18

- Objective:
  - Select the preferences relevant for the current context
- Input:
  - The current context
  - The preference profile
- Output:
  - The list of active preferences



Preference Selection – Relevance relation

19

- **Relevance**:
  - A partial order relationship (abstractness) is defined among context configurations (CDT hierarchical structure)
  - A preference is relevant for the current context if its context configuration is equal or more abstract than the current context
- **Distance index**:

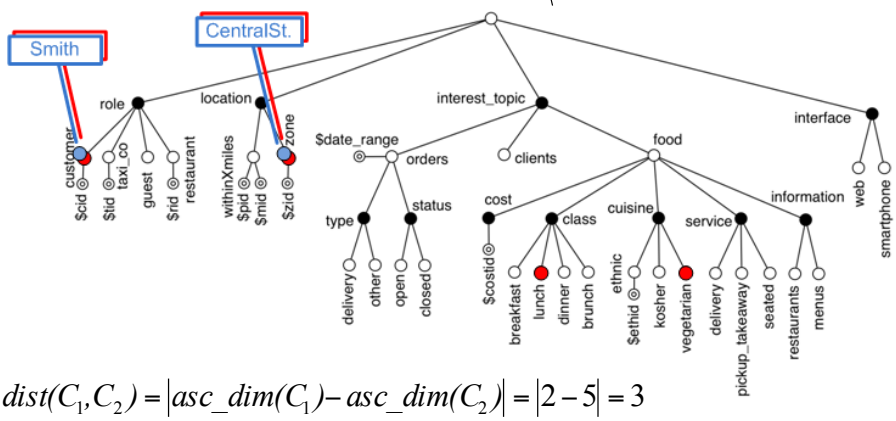
$$dist(C_1, C_2) = |asc\_dim(C_1) - asc\_dim(C_2)|$$

  - *asc_dim(C_i)* = number of dimension nodes that are ancestor of the instantiated context value nodes of C_i
  - Represents the number of dimension nodes present on the CDT between the two context configurations

Preference selection – Relevance relation Example

20

$$C_1 = \left\langle \begin{array}{l} role = client(\text{"Smith"}) \wedge \\ location = zone(\text{"CentralSt."}) \end{array} \right\rangle \succ C_2 = \left\langle \begin{array}{l} role = client(\text{"Smith"}) \wedge \\ location = zone(\text{"CentralSt."}) \wedge \\ cuisine = vegetarian \wedge \\ class = lunch \end{array} \right\rangle$$



$$dist(C_1, C_2) = |asc\_dim(C_1) - asc\_dim(C_2)| = |2 - 5| = 3$$

Preference selection - Steps

21

- Select from the context profile preferences relevant for the current context
- Assign to each relevant preference *cp* a **relevance index**:

$$relevance(cp) = \frac{dist(C_{curr\_context}, C_{root\_context}) - dist(C_{cp}, C_{curr\_context})}{dist(C_{curr\_context}, C_{root\_context})}$$

  - It is a relevance percentage w.r.t. the current context
  - *Relevance = 1* for preferences expressed on the current context
  - *Relevance = 0* for preferences expressed on the root context

Preference selection – Example

22

Current Context: $\boxed{dist(c_{curr}, c_{root}) = 4}$

$C_{curr} = \langle \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \wedge \texttt{information} = \texttt{restaurants} \rangle$

Preference Profile:

$\boxed{dist(c_{curr}, c_1) = 0 \quad relevance = 1}$

$$\begin{aligned} CP_1 = \langle C_1 &= \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \\ &\quad \wedge \texttt{information} = \texttt{restaurant}, \\ SQ_{\sigma 1} &= \texttt{restaurant} \ltimes \texttt{restaurant\_cuisine} \ltimes \sigma_{\texttt{cuisine.description} = \text{“Chinese”}} \texttt{cuisine}, \\ S_{\sigma 1} &= 0.8 \rangle \end{aligned}$$

$\boxed{dist(c_{curr}, c_2) = 1 \quad relevance = 0.75}$

$$\begin{aligned} CP_2 = \langle C_2 &= \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{information} = \texttt{restaurant}, \\ SQ_{\sigma 2} &= \texttt{restaurant} \ltimes \texttt{restaurant\_cuisine} \ltimes \sigma_{\texttt{cuisine.description} = \text{“Chinese”}} \texttt{cuisine}, \\ S_{\sigma 2} &= 0.5 \rangle \end{aligned}$$

$\boxed{dist(c_{curr}, c_3) = 3 \quad relevance = 0.25}$

$$\begin{aligned} CP_3 = \langle C_3 &= \texttt{role} = \texttt{client}(\text{“Smith”}), \\ SQ_{\sigma 3} &= \sigma_{\texttt{isVegeterian}=1}(\texttt{dishes}), \\ S_{\sigma 3} &= 0.3 \rangle \end{aligned}$$

$\boxed{dist(c_{curr}, c_4) = 3 \quad relevance = 0.25}$

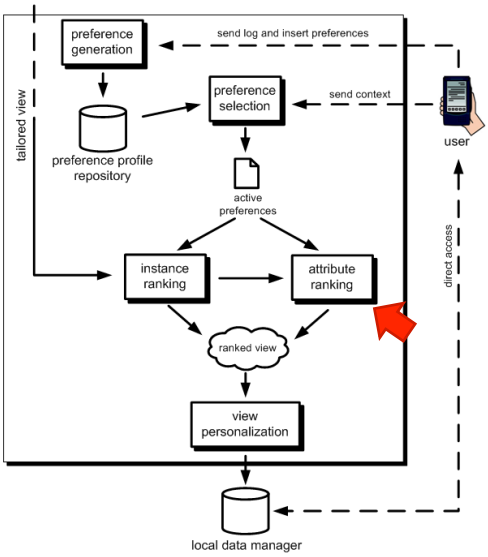
$$\begin{aligned} CP_4 = \langle C_4 &= \texttt{role} = \texttt{client}(\text{“Smith”}), \\ SQ_{\sigma 4} &= \sigma_{\texttt{isMildSpicy}=1}(\texttt{dishes}), \\ S_{\sigma 4} &= 0.8 \rangle \end{aligned}$$

$$\begin{aligned} CP_5 = \langle C_5 &= \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \\ &\quad \wedge \texttt{interface} = \texttt{smartphone}, \\ A_{\pi 5} &= \{\texttt{name}, \texttt{zipcode}, \texttt{phone}\}, \\ S_{\pi 5} &= 0.8 \rangle \end{aligned}$$

$$\begin{aligned} CP_6 = \langle C_6 &= \texttt{role} = \texttt{client}(\text{“Smith”}) \wedge \texttt{location} = \texttt{zone}(\text{“CentralSt.”}) \\ &\quad \wedge \texttt{interface} = \texttt{smartphone}, \\ A_{\pi 6} &= \{\texttt{address}, \texttt{city}\}, \\ S_{\pi 6} &= 0.2 \rangle \end{aligned}$$

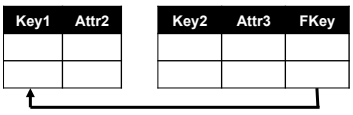
Attribute Ranking – Overview

- Objective:
  - Score attributes of the view tailored by the designer
- Input:
  - The schema of the tailored view
  - The list of active π-preferences
- Output:
  - The ranked schema of the tailored view



Attribute ranking – Steps

- Select for each attribute the preferences related to it
- Combine scores related to the same attribute, exploiting the relevance
- Rank attributes on which no preference is expressed with an indifference score (0.5)
- Satisfy referential integrity constraints
  - Primary keys and foreign keys must have the highest score in the table
  - Attributes referenced by foreign keys must have a score higher than the related foreign keys



$$Key1.score \leftarrow max(Key1.score, Attr2.score)$$
$$Key2.score \leftarrow max(Key2.score, Attr3.score, FKey.score)$$
$$FKey.score \leftarrow max(Key2.score, Attr3.score, FKey.score)$$
$$Key1.score \leftarrow max(Key1.score, FKey.score)$$

Attribute ranking – Example

25

RESTAURANTS(restaurant_id, name, address, zipcode, city, phone, fax, email, website, openinghourslunch, openinghoursdinner, closingday, capacity, parking)
RESTAURANT_CUISINE(restaurant_id, cuisine_id)
CUISINES(cuisine_id, description)

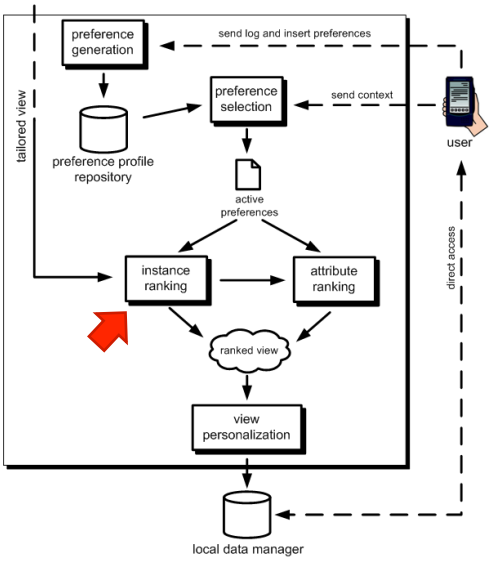
$$P_{\pi 1} = \langle \{\texttt{name}, \texttt{cuisine.description}, \texttt{phone}, \texttt{closingday}\}, 1 \rangle, R = 1$$
$$P_{\pi 2} = \langle \{\texttt{address}, \texttt{city}, \texttt{state}, \texttt{phone}\}, 0.1 \rangle, R = 0.2$$
$$P_{\pi 3} = \langle \{\texttt{fax}, \texttt{email}, \texttt{website}\}, 0.1 \rangle, R = 0.2$$

RESTAURANTS(restaurant_id:1, name:1, address:0.1, zipcode:0.5, city:0.1, phone:1, fax:0.1, email:0.1, website:0.1, openinghourslunch:0.5, openinghoursdinner:0.5, closingday:1, capacity:0.5, parking:0.5)
RESTAURANT_CUISINE(restaurant_id:0.5, cuisine_id:0.5)
CUISINES(cuisine_id:1, description:1)

Tuple ranking – Overview

26

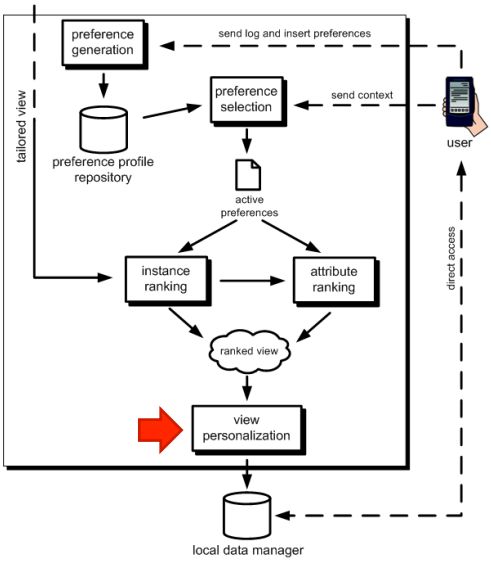
- Objective:
  - Score tuples of the view tailored by the designer
- Input:
  - The tailored view
  - The list of active σ- preferences
- Output:
  - The tailored view with ranked data tuples



View personalization – Overview

27

- Objective:
  - Customize the tailored view
  - Fulfill memory space constraints
- Input:
  - The view with ranked attributes and tuples
- Output:
  - The customized view



View personalization – Steps

28

- The view personalization process can only reduce the view proposed by the designer
- View personalization strategy
  1. Filtering attributes on the basis of a score threshold
  2. *Top-K* selection on each table
- *K* computed as a function of
  - The average table attribute score w.r.t. the overall view attribute score
  - Memory occupation of the table

View personalization - Example

RESTAURANTS(restaurant_id:1, name:1, address:0.1, zipcode:0.5, city:0.1, phone:1, fax:0.1, email:0.1, website:0.1, openinghourslunch:0.5, openinghoursdinner:0.5, closingday:1, capacity:0.5, parking:0.5)
RESTAURANT_CUISINE(restaurant_id:0.5, cuisine_id:0.5)
CUISINES(cuisine_id:1, description:1)

threshold = 0.5

RESTAURANTS(restaurant_id:1, name:1, zipcode:0.5, phone:1, openinghourslunch:0.5, openinghoursdinner:0.5, closingday:1, capacity:0.5, parking:0.5)
RESTAURANT_CUISINE(restaurant_id:0.5, cuisine_id:0.5)
CUISINES(cuisine_id:1, description:1)

View personalization – Example (2)

RESTAURANTS(restaurant_id:1, name:1, zipcode:0.5, phone:1, openinghourslunch:0.5, openinghoursdinner:0.5, closingday:1, capacity:0.5, parking:0.5)
RESTAURANT_CUISINE(restaurant_id:0.5, cuisine_id:0.5)
CUISINES(cuisine_id:1, description:1)

compute average schema score

Table	Average Score
RESTAURANTS	0.72
RESTAURANT_CUISINE	0.5
CUISINES	1
RESTAURANT_SERVICE	0.5
SERVICE	0.6
RESERVATION	0.72

View personalization – Example (3)

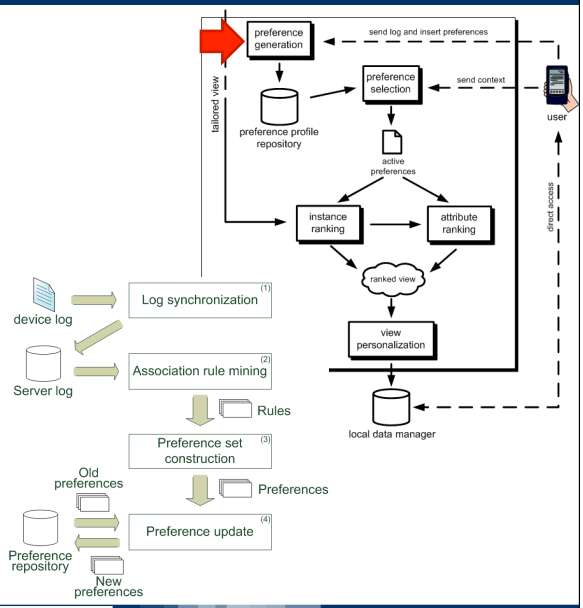
Table	Average Score
RESTAURANTS	0.72
RESTAURANT_CUISINE	0.5
CUISINES	1
RESTAURANT_SERVICE	0.5
SERVICE	0.6
RESERVATION	0.72

order tables and partition available space (2Mb)

Table	Average Score	Memory (Mb)
CUISINES	1	0.50
RESTAURANTS	0.72	0.35
RESERVATION	0.72	0.35
SERVICE	0.6	0.30
RESTAURANT_CUISINE	0.5	0.25
RESTAURANT_SERVICE	0.5	0.25

Preference generation – Overview

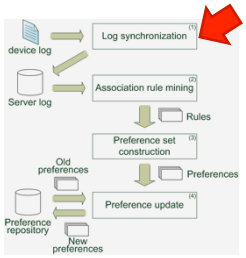
- Objective:
  - Generate preference profiles analyzing log data, extracting knowledge in terms of association rules
- Input:
  - User activity log
  - Old preferences
- Output:
  - σ- and π-preferences
- The sub-tasks are performed independently for σ- and π-preferences



Mining σ-preferences
Log synchronization

33

- **Input**: *log stored on the client device*
  - Textual recording of SQL queries



- **Output**: *log on the server*
  - A relational table for each table in the database
  - The log associated with a table R(X) contains:
    - An attribute for each black node of the CDT
    - An attribute for each attribute belonging to X
    - An attribute for each attribute of each table reachable from R(X) through foreign key constraints
  - A row for each tuple returned in user query answers

POLITECNICO DI MILANO

Mining σ-preferences
Server log – Example

34

SELECT DISTINCT dishes.description
FROM dishes, restaurants, restaurant_dish
WHERE restaurants.restaurant_id= restaurant_dish.restaurant_id
AND dishes.dish_id= restaurant_dish.dish_id
AND restaurants.closingday='Monday'

RESTAURANTS

restaurant_id	name	closingday
r1	Pizzeria Rita	Monday
r2	Cing Restaurant	Tuesday
r3	Cantina Mariachi	Monday

DISHES

dish_id	description
p1	Pizza Margherita
p2	Pizza Napoli
p3	Pizza Capricciosa

RESTAURANT_DISH

restaurant_id	dish_id
r1	p1
r1	p3
r2	p2
r2	p3
r3	p1

Log of the table Restaurant_dish

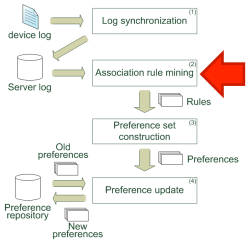
Context dimensions					Restaurant_dish attributes		Restaurants attributes			Dishes attributes	
id	role	int-topic	cuisine	...	rd.r_id	rd.d_id	r.r_id	r.name	r.closingday	d.d_id	d.descr
1	client(cli1)	food	veg		r1	p1	r1	Pizzeria Rita	Monday	p1	Margh.
2	client(cli1)	food	veg		r1	p3	r1	Pizzeria Rita	Monday	p3	Capr.
3	client(cli1)	food	veg		r3	p1	r3	Cantina Mariachi	Monday	p1	Margh.

POLITECNICO DI MILANO

Mining σ-preferences
Association rule mining

35

- **Input**: *server log*
- **Output**: *association rules*



- Knowledge is extracted from the log by means of association rules
- An association rule is an implication in the form

$$A \rightarrow B$$

- Quality indexes for association rules:

$$\text{Support} = \text{num. of data with } A \cup B$$

$$\text{Confidence} = \frac{\text{num. of data with } A \cup B}{\text{num. of data with } A}$$

POLITECNICO DI MILANO

Mining σ-preferences
Association rule mining (2)

36

- We are interested in **σ-rules**, correlating contexts and data
- A σ-rule on a relation R(X) is a triple:

$$\langle C \rightarrow cond, sup, conf \rangle$$

- *C*: a context
- *Cond:* a conjunction of conditions in the form *A=value*, where *A* is an attribute belonging to R(X) or to a relation reachable from R(X) through foreign keys
- *sup:* the support of the association rule C → cond
- *conf*: is the confidence of the association rule C → cond

- Example:

$$\langle \text{location} = \text{zone}(\text{"Central St."}) \wedge \text{interest-topic} = \text{food} \rightarrow \text{isVegetarian} = \text{true} \wedge \text{isSpicy} = \text{false}, 0.2, 0.7 \rangle$$

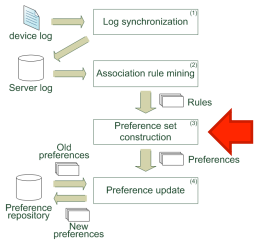
POLITECNICO DI MILANO

Mining σ-preferences
Preference-set construction

37

- **Input**: *σ-rules*
- **Output**: *σ-preferences*

- Three sub-tasks:
  - ***Removal of redundant rules***, whose satisfaction is implied by other ones
  - ***Computation of the score*** of the σ-preferences



Mining σ-preferences
Score computation

38

- A σ-rule *r* = <*C* → *cond, sup, conf*> may determine a σ-preference *p* = <*C*, <SQ_σ, *score*>> such that *p.C* = *r.c* and $p.SQ_\sigma$ = *C.cond*
- The mining procedure can discover only preferences indicating user interest in the interval (0.5, 1], and *only if the confidence of the rule is greater than the frequency of the data satisfying it*

- The score is computed as follows:

$$\Delta = r.conf - f$$

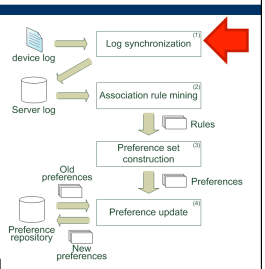
$$p.score = \min\left((1+\gamma\cdot\Delta)\cdot 0.5,\ 1\right),\ \text{if } \Delta > 0$$

- *f* is the frequency of the data satisfying *r.cond* in the data set accessed by the user

Mining π-preferences
Log synchronization

39

- **Input**: *log stored on the client device*
  - Textual recording of SQL queries
- **Output**: *log on the server*
  - An attribute for each black node of the CDT
  - An attribute for each attribute of each table
  - A row for each query
  - Non-contextual columns contain '1' if the associated attribute has been accessed in the query
  - A query accesses an attribute if it is contained in the select or where clause




```
SELECT restaurants.address,
       restaurants.phone
FROM restaurants
WHERE restaurants.closingday<>'lunedì'
```

```
SELECT dishes.description
FROM dishes
WHERE dishes.category_id='high'
```

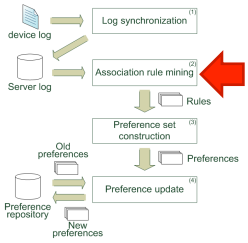
Context dimensions — Attributes of the database relations

role	interest-topic	cuisine	r.rest_id	r.addr	r.city	r.phone	r.closingday	d.dish_id	d.descr	d.isSpicy	d.isMildSpicy	d.cat_id
client(Smith)	food	veget		1		1	1					
client(Smith)	food	veget							1			1

Mining π-preferences
Association rule mining

40

- **Input**: *server log*
- **Output**: *association rules*



- We are interested in π-rules, correlating contexts and attributes
- A **π-rule** on a relation R(X) is a triple:

  <C → Attr, sup, conf>
- *C*: a context
- *Attr:* an attribute
- *sup:* is the support of the association rule *C → Attr*
- *conf:* is the confidence of the association rule *C → Attr*

- Example:

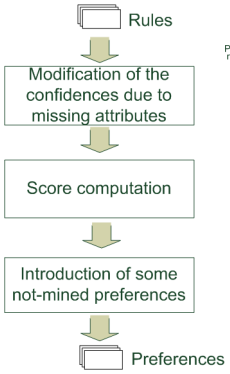
$$< \text{interest-topic} = \text{food} \rightarrow \text{isVegetarian}, 0.1, 0.7 >$$

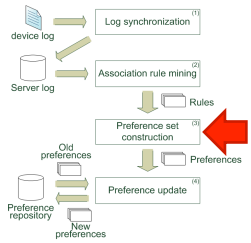
Mining π-preferences
Preference-set construction

41

- **Input**: *π-rules*
- **Output**: *π-preferences*

- Three sub-tasks:
  - ***Modification of the confidences of some rules*** due to missing attributes
  - ***Computation of the score*** of the π-preferences
  - ***Introduction of some not-mined preferences*** to deal with preference-propagation problems

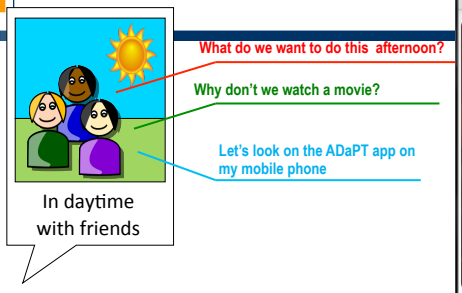


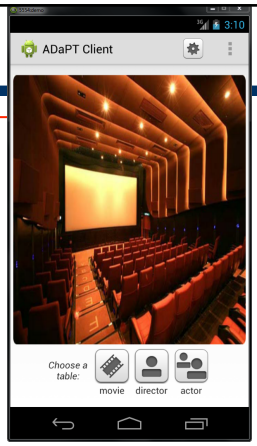


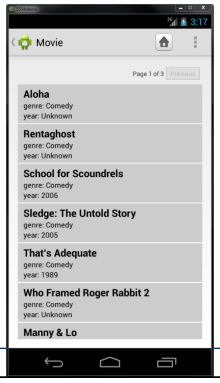
POLITECNICO DI MILANO

42

ADaPT: ~~Automatic Data~~ Personalization Based on Contextual Preferences [ICDE 2014]







Movie list has changed!