# Recommending Multidimensional Queries

Arnaud Giacometti, Patrick Marcel, Elsa Negre

LI - Université François Rabelais Tours

DAWAK'09

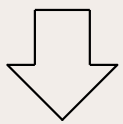
# Outline

# Motivations & Intuitions

## The problem

Navigate an OLAP cube:

- an analysis session
- the forthcoming query?



How to propose to the user his forthcoming query ?

# Motivations & Intuitions

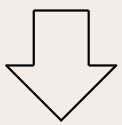
## Related work

- [Khoussainova *et al.*, CIDR09]: DB & recommendations?
- [Chatzopoulou G., SSDBM09]: DB recommending SQL queries
- [Sarawagi, VLDB00, Sarawagi and Sathe, SIGMOD00, Cariou *et al.*, DaWaK08]: Discovery driven analysis
- [Jerbi *et al.*, DaWaK09]: Content-based filtering

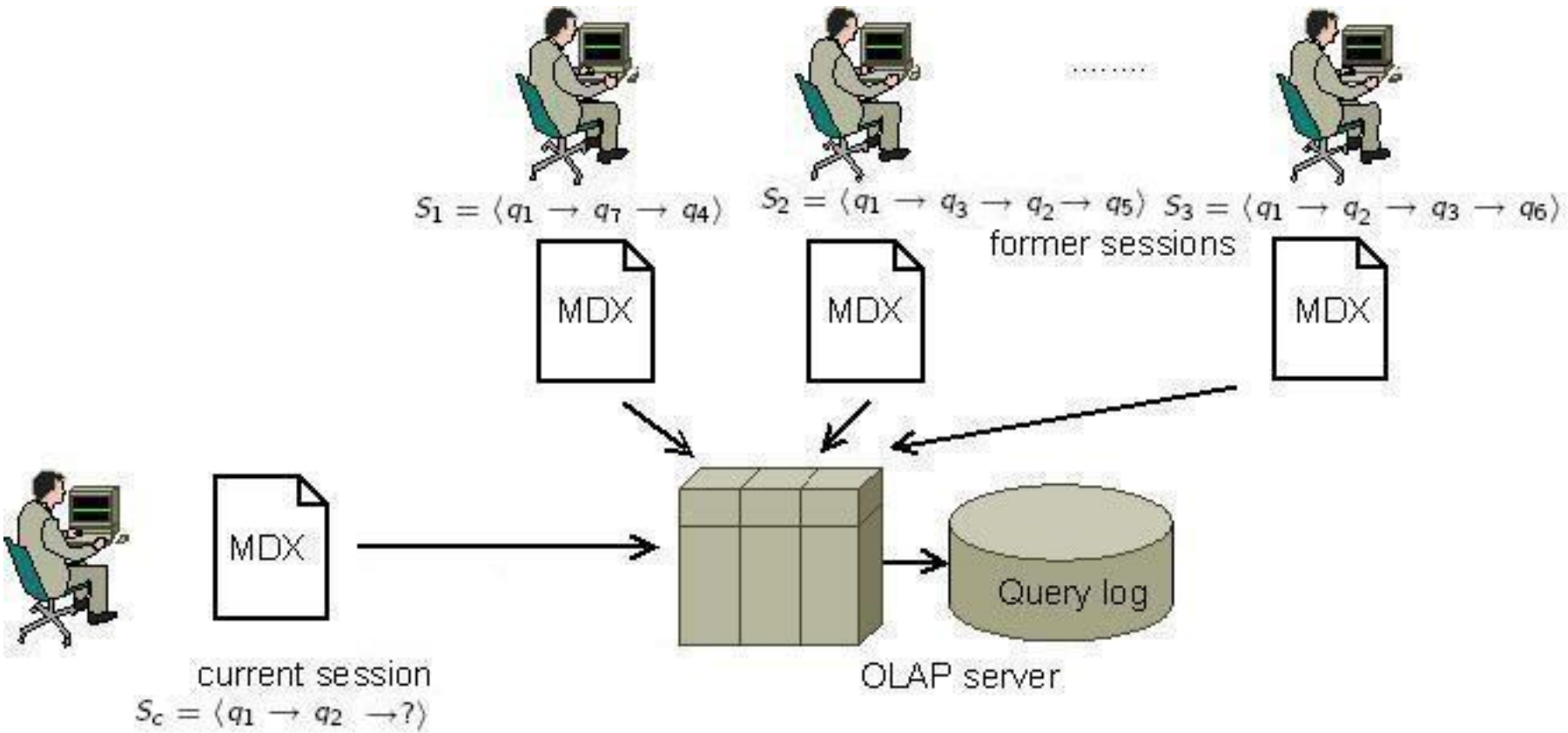
# Motivations & Intuitions

## Proposal

Collaborative filtering in:

| Information Retrieval<br>Web Usage Mining | OLAP |
|---|---|
|  |  |
| Use the known behaviors of a population<br>to envisage the future actions<br>of a particular user<br>and Seek, by comparison,<br>the users having similar behaviors | Exploitation of<br>the other users former navigations<br>to generate recommendations |

# Logs


$S_1 = \langle q_1 \rightarrow q_7 \rightarrow q_4 \rangle$
$S_2 = \langle q_1 \rightarrow q_3 \rightarrow q_2 \rightarrow q_5 \rangle$
$S_3 = \langle q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_6 \rangle$
former sessions
MDX
MDX
MDX
MDX
Query log
current session
$S_c = \langle q_1 \rightarrow q_2 \rightarrow ? \rangle$
OLAP server

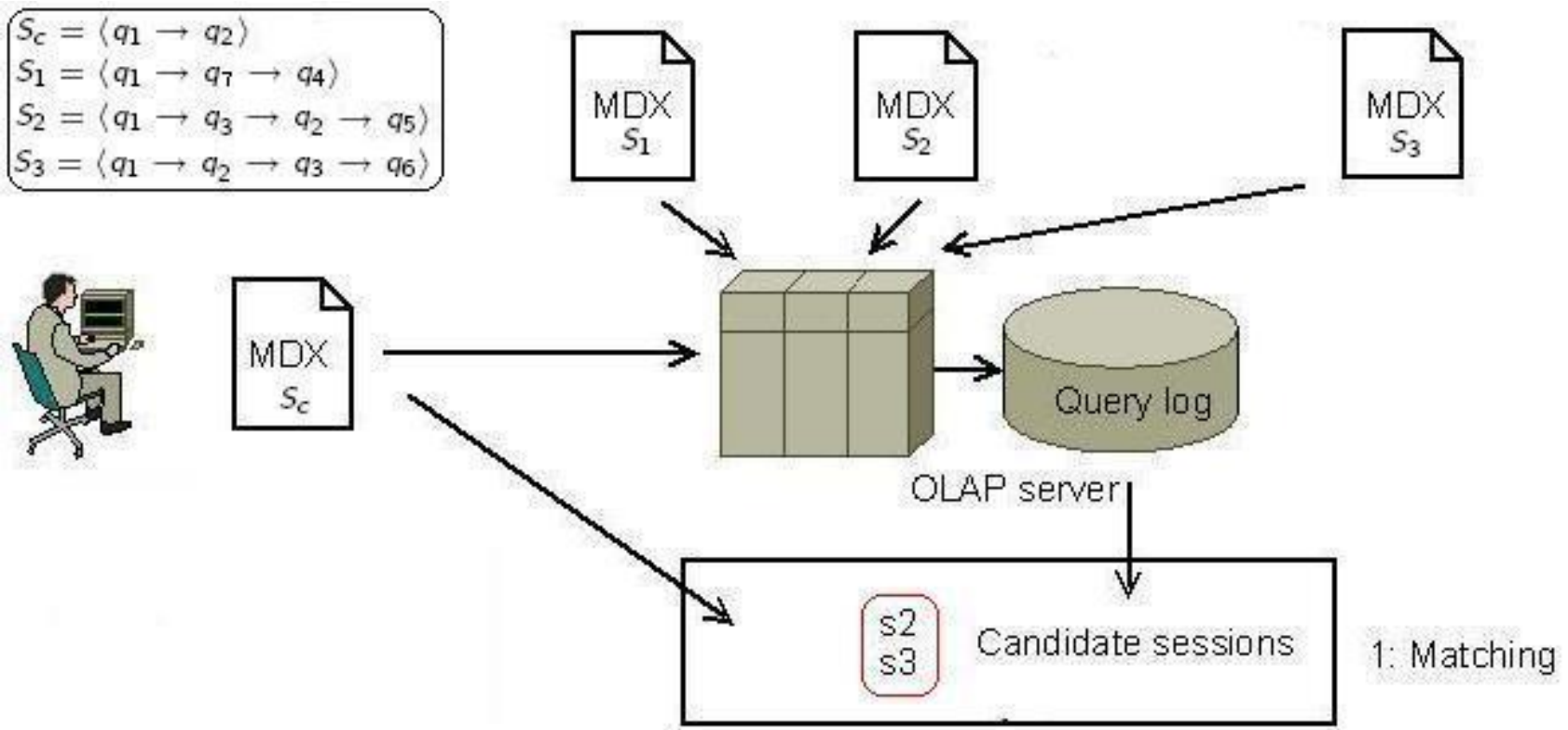
# Step 1: Matching

Obtain sessions matching the current session

**How?**

**Candidate Sessions**: Comparison of queries sequences (sessions)
**Need**: Distance between sessions

# Step 1: Matching

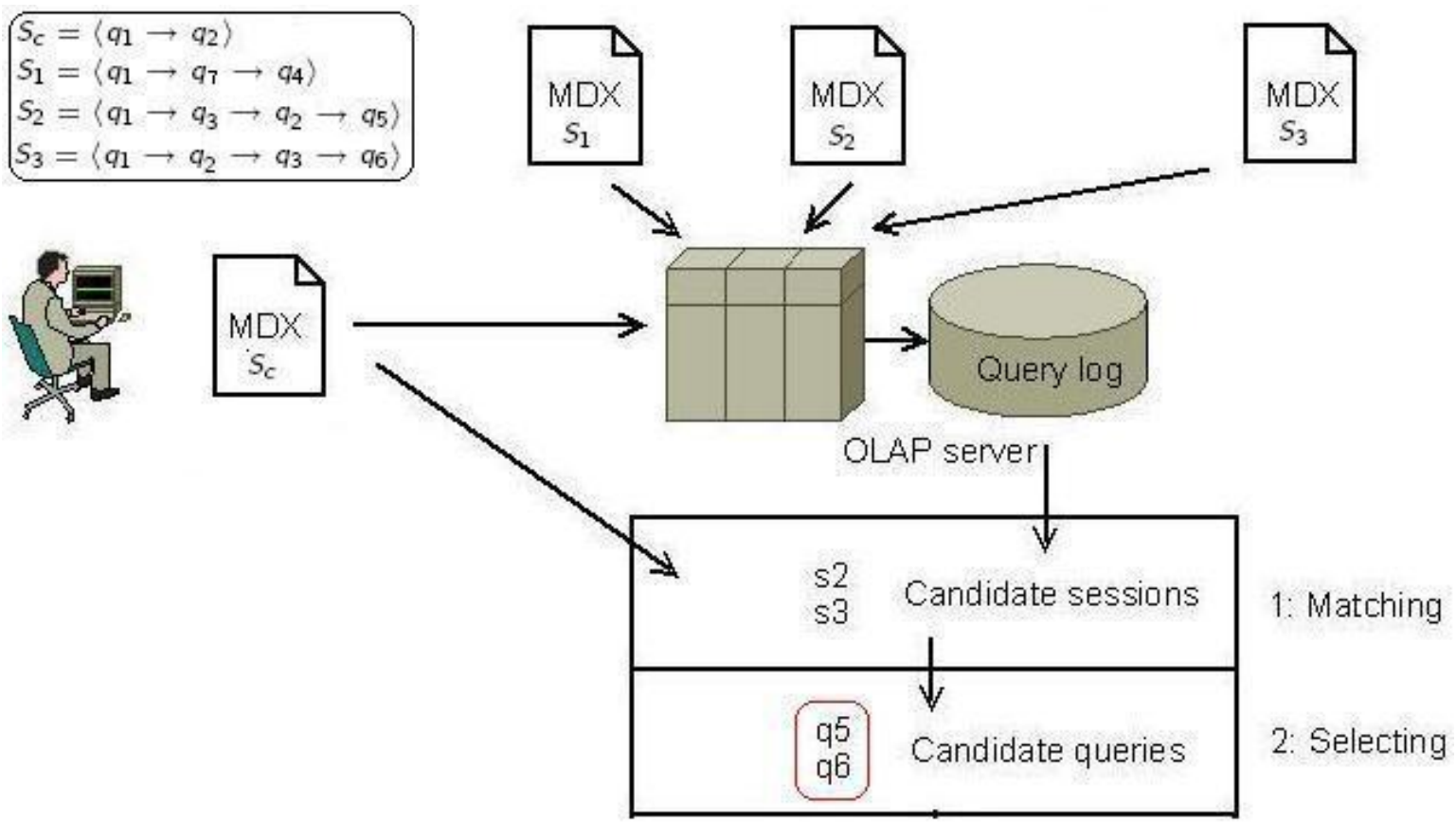


# Step 2: Selecting

## Select queries contained in candidate sessions

### How?

**Candidate Queries**: the last query of each candidate session
Analogy with Web : the last query contains the goal of the session



# Step 3: Ranking

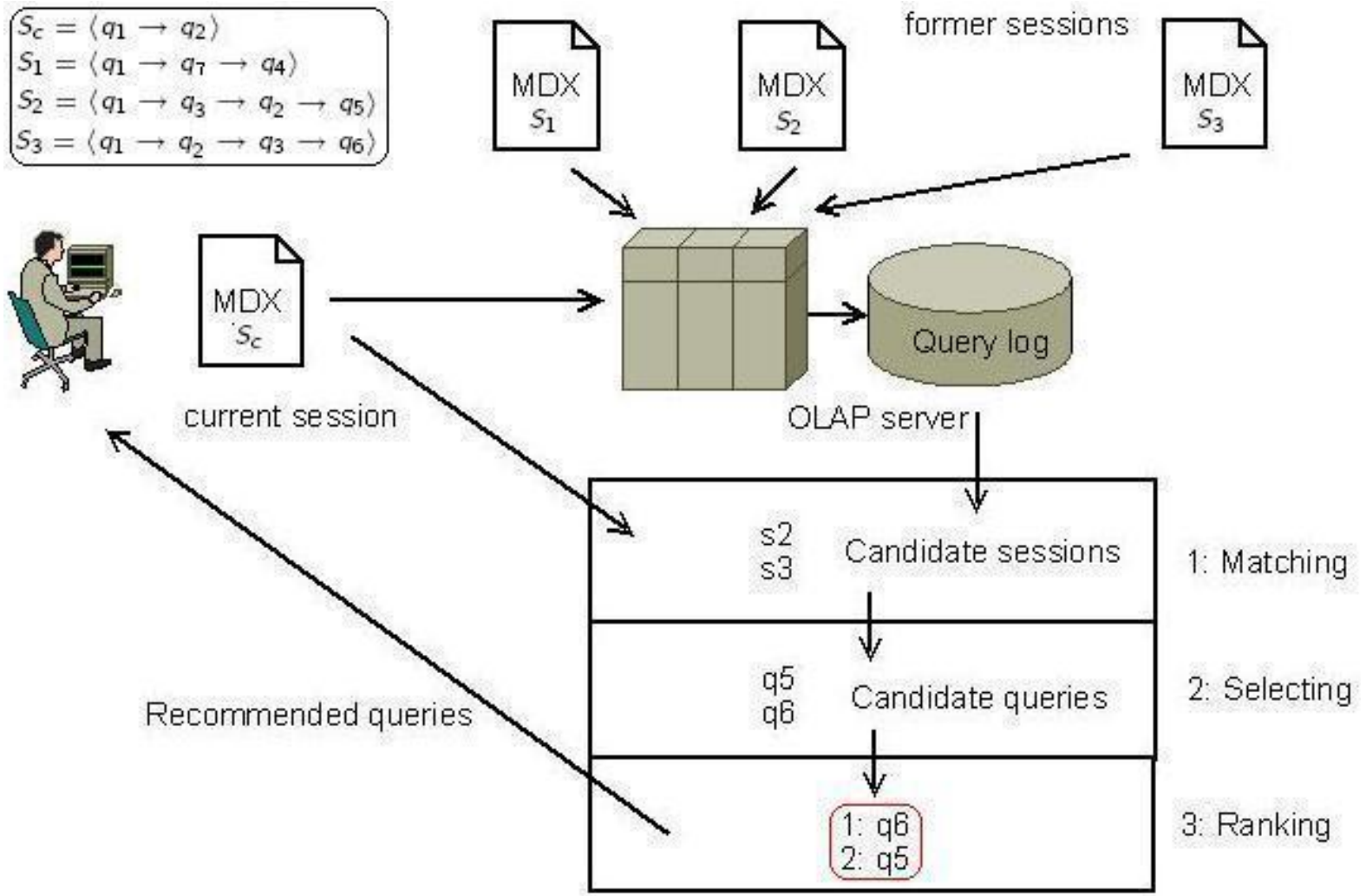
## Ranking candidate queries

**How?**

**Recommended Queries**: the closest to the last query of the current session in the sense of the distance between queries.

**Need**: Distance between Queries - Hausdorff distance

# Step 3: Ranking



# Definitions by the practice

## Distance between References

- Queries:

# Definitions by the practice

## Distance between References

- Queries:
  - $q_1$: Number of sales of drinks in *France* in 2007 and 2008:
    $\{\langle Drink, France, 2007\rangle, \langle Drink, France, 2008\rangle\} = \{r_1^1, r_1^2\}$

| SALES : Number | | France |
|---|---|---|
| Drink | 2007 | 10 |
| | 2008 | 20 |

# Definitions by the practice

## Distance between References

- Queries:
  - $q_1$: Number of sales of drinks in *France* in 2007 and 2008:
    $\{\langle Drink, France, 2007\rangle, \langle Drink, France, 2008\rangle\} = \{r_1^1, r_1^2\}$

| SALES : Number | | France |
|---|---|---|
| Drink | 2007 | 10 |
| | 2008 | 20 |

  - $q_2$: Number of sales of all products in *Austria* in 2008: $\{\langle All, Austria, 2008\rangle\} = \{r_2^1\}$

| SALES : Number | | Austria |
|---|---|---|
| All | 2008 | 15 |

# Definitions by the practice

## Distance between References

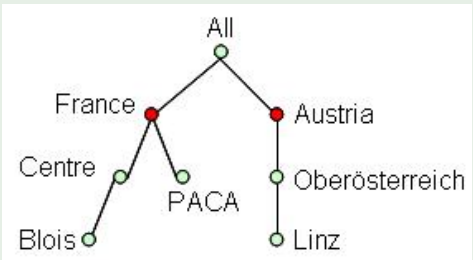
- Queries:
  - $q_1$: Number of sales of drinks in *France* in 2007 and 2008:
    $\{\langle Drink, France, 2007\rangle, \langle Drink, France, 2008\rangle\} = \{r_1^1, r_1^2\}$

| SALES : Number | | France |
|---|---|---|
| Drink | 2007 | 10 |
| | 2008 | 20 |

  - $q_2$: Number of sales of all products in *Austria* in 2008: $\{\langle All, Austria, 2008\rangle\} = \{r_2^1\}$

| SALES : Number | | Austria |
|---|---|---|
| All | 2008 | 15 |

- Distance between members: shortest path



# Definitions by the practice

## Distance between References

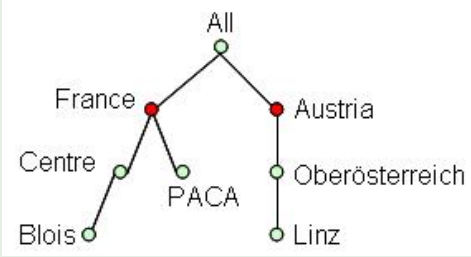
- Queries:
  - $q_1$: Number of sales of drinks in *France* in 2007 and 2008:
    $\{\langle Drink, France, 2007\rangle, \langle Drink, France, 2008\rangle\} = \{r_1^1, r_1^2\}$

| SALES : Number | | France |
|---|---|---|
| Drink | 2007 | 10 |
| | 2008 | 20 |

  - $q_2$: Number of sales of all products in *Austria* in 2008: $\{\langle All, Austria, 2008\rangle\} = \{r_2^1\}$

| SALES : Number | | Austria |
|---|---|---|
| All | 2008 | 15 |

- Distance between members: shortest path




- Distance between references:

$$d_{ref}(r_1^2, r_2^1) = d_{members}(Drink, All) + d_{members}(France, Austria) + d_{members}(2008, 2008)$$
$$= 2 + 2 + 0 = 4$$

# Definitions by the practice

## Distance between Queries

- Queries:
  - $q_1$: Number of sales of drinks in *France* in 2007 and 2008:
    $\{\langle Drink, France, 2007\rangle, \langle Drink, France, 2008\rangle\} = \{r_1^1, r_1^2\}$

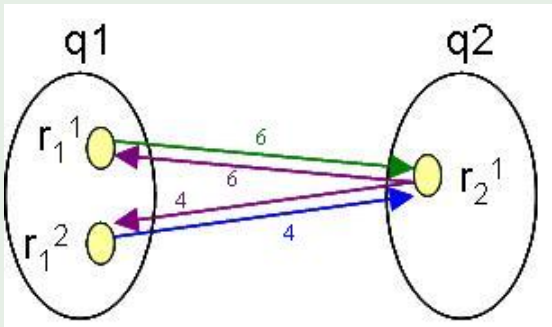
| SALES : Number | | France |
|---|---|---|
| Drink | 2007 | 10 |
| | 2008 | 20 |

  - $q_2$: Number of sales of all products in *Austria* in 2008: $\{\langle All, Austria, 2008\rangle\} = \{r_2^1\}$

| SALES : Number | | Austria |
|---|---|---|
| All | 2008 | 15 |

- Hausdorff distance:

$$d_h(q_1, q_2) = \max\{\, \max_{r_1 \in q_1} \min_{r_2 \in q_2} d_{ref}(r_1, r_2),$$
$$\max_{r_2 \in q_2} \min_{r_1 \in q_1} d_{ref}(r_1, r_2)\,\} = 6$$



# Definitions by the practice

## Distance between Sequences

### Words (sequences of letters)

$$d_{ed}(CAR, CAT)$$

Operations costs :

substitution of a letter by another = 1

insertion (or deletion) of a letter = 1

| | | C | A | R |
|---|---|---|---|---|
| | 0 | 1 | 2 | 3 |
| C | 1 | 0 | 1 | 2 |
| A | 2 | 1 | 0 | 1 |
| T | 3 | 2 | 1 | 1 |

$$CAR \overset{subst(R,T)}{\rightarrow} CAT = 1$$

### Sessions (sequences of queries)

$$d_{ed}(q_3, q_1 \rightarrow q_2)$$

Operations costs :

substitution of a query q by another q'

$$= d_H(q, q')$$

insertion (or deletion) of a query = 2

| | | q1 | q2 |
|---|---|---|---|
| | 0 | 2 | 4 |
| q3 | 2 | dH(q1,q3) | dH(q1,q3) + 2 |

$$q_3 \overset{subst(q_3,q_1)}{\rightarrow} q_1 \overset{add(q_2)}{\rightarrow} (q_1 \rightarrow q_2)$$

$$= d_H(q_1, q_3) + 2$$

# Presentation

## Our generator

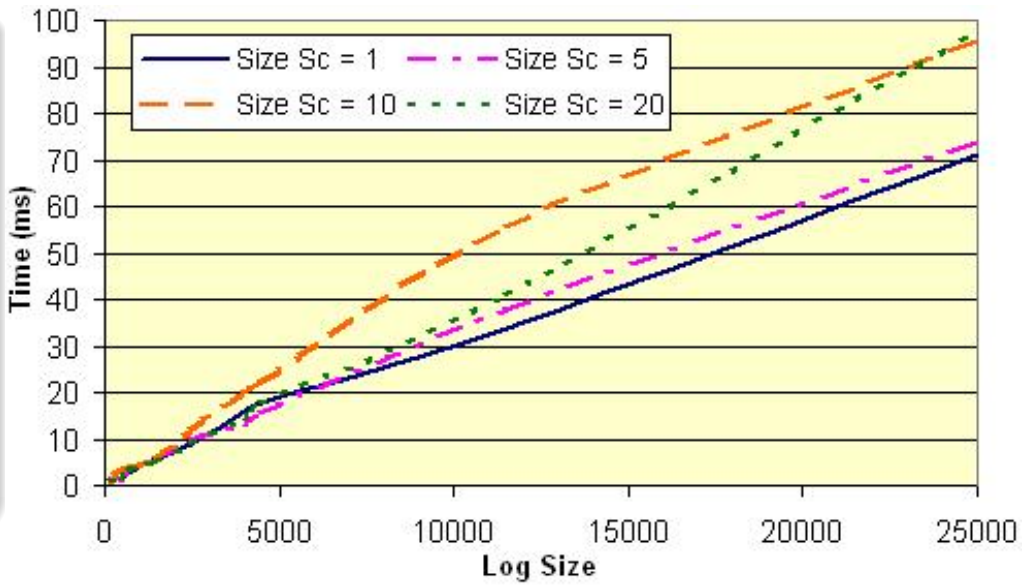
- The cube
  - Foodmart database (Mondrian OLAP engine)
- The sessions
  - max 100 references per MDX query
  - X sessions in log
  - max Y queries per session
- Property
  - high density of the generated log

# Performance analysis

Measure of the time taken to propose a recommendation

## Protocol

- 25<X<500 sessions
- 20<Y<50 queries per session
- 150<log<25 000 queries
- Current session : 1<Y<20 queries




## Observations

- Time increases slowly with log size
- Negligible time (<100ms)

# Precision/Recall analysis

## Protocol

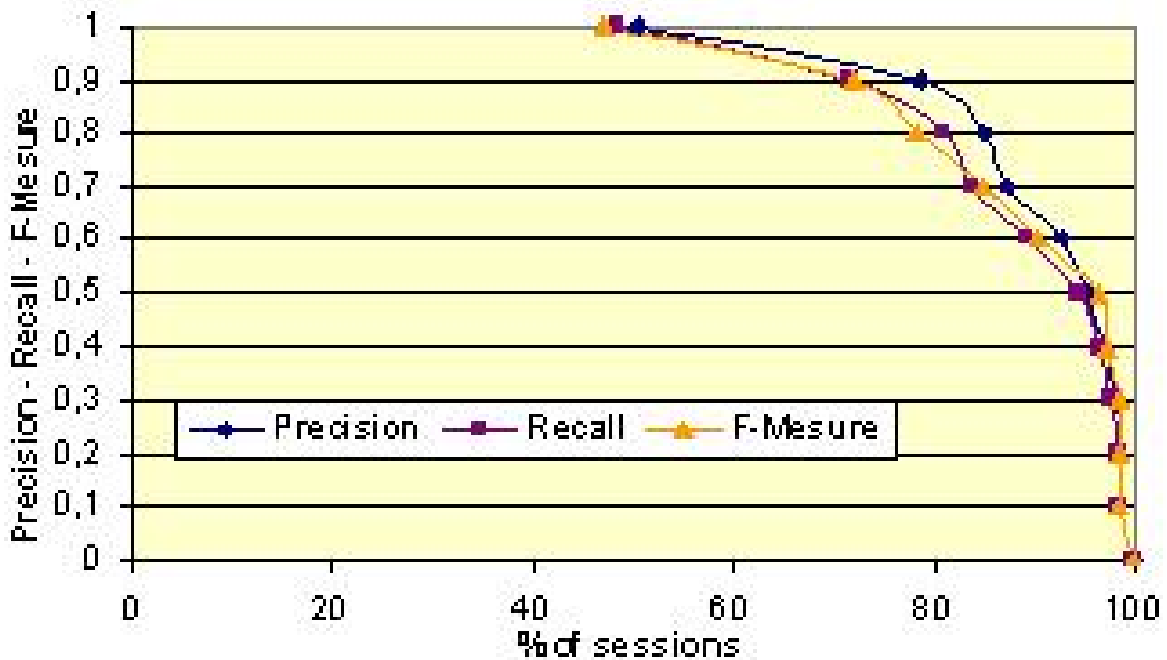
- 10-fold cross validation
  $\Rightarrow$ Assess how the results of our proposition will generalize to an independent data set [Chatzopoulou G., SSDBM09].
  - Generated set of sessions : 10 equally sized subsets
  - Log = 9 subsets
  - Current session = each session of 1 subset without the last query
  - Last query of current session : the expected query ($q_{ex}$)
  - Computation of the recommended query ($q_{rec}$) for $q_{ex}$

# Precision/Recall analysis

Is the recommended query the expected query?

## Test

*inverse CFD* [Precision, Recall, F-measure]




## Observations

Effectiveness of the proposition for dense log

# Conclusion & Future work

## Contributions

- A method to propose MDX queries as recommendations
- Experiments results:
  - Recommendations can be computed efficiently
  - Objectively good recommendations

## Future work

- Query recommendations for OLAP discovery driven analysis [Giacometti *et al.*, DOLAP09]
- Incorporate OLAP query personalization
- Experiments on real data with users feedbacks
- Contributing to a collaborative query management system [Khoussainova *et al.*, CIDR09]

Thank you for your attention.
Any questions?

# References I


Véronique Cariou, Jérôme Cubillé, Christian Derquenne, Sabine Goutier, Françoise Guisnel, and Henri Klajnmic.
Built-in indicators to discover interesting drill paths in a cube.
In *DaWaK '08: Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, pages 33–44, Berlin, Heidelberg, DaWaK'08. Springer-Verlag.


Polyzotis N. Chatzopoulou G., Eirinaki M.
Query recommendations for interactive database exploration.
In *21st International Conference on Scientific and Statistical Database Management,June 2-4, 2009, New Orleans, Louisiana USA*, SSDBM'09.

# References II


Arnaud Giacometti, Patrick Marcel, Elsa Negre, and Arnaud Soulet.
Query recommendations for OLAP discovery driven analysis.
DOLAP'09.

Houssem Jerbi, Franck Ravat, Olivier Teste, and Gilles Zurfluh.
Preference-based recommendations for OLAP analysis.
DaWaK'09.

Nodira Khoussainova, Magdalena Balazinska, Wolfgang Gatterbauer, YongChul Kwon, and Dan Suciu.
A case for a collaborative query management system.
www.crdrdb.org, CIDR'09.

# References III


Sunita Sarawagi and Gayatri Sathe.
i$^3$: Intelligent, interactive investigaton of olap data cubes.
In *SIGMOD Conference*, page 589, SIGMOD'00.


Sunita Sarawagi.
User-adaptive exploration of multidimensional data.
In *VLDB*, pages 307–316, VLDB'00.

## Precision / Recall / F-measure

- Precision $= \frac{|members(q_{ex}) \cap members(q_{rec})|}{|members(q_{rec})|}$
- Recall $= \frac{|members(q_{ex}) \cap members(q_{rec})|}{|members(q_{ex})|}$
- F-measure $= \frac{2 \times (precision \times recall)}{(precision + recall)}$