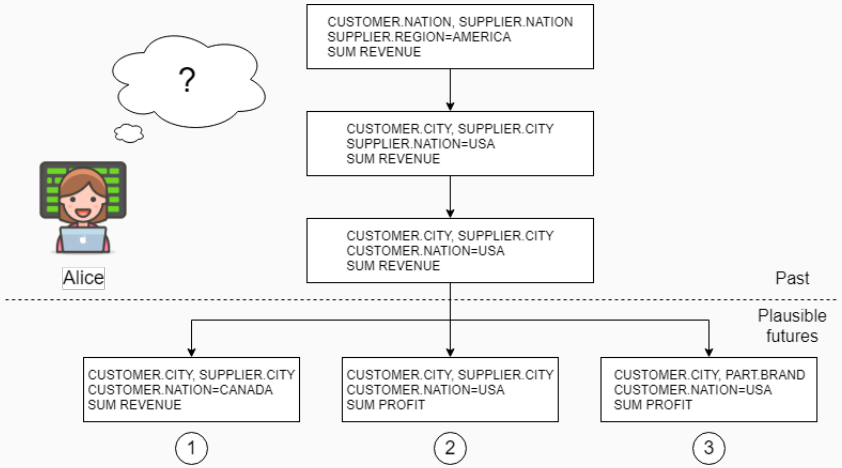
Profiling user belief in BI exploration for measuring subjective interestingness

Alexandre Chanson, Ben Crulis, Krista Drushku,
Nicolas Labroche, Patrick Marcel

DOLAP 2019 - 26 March 2019

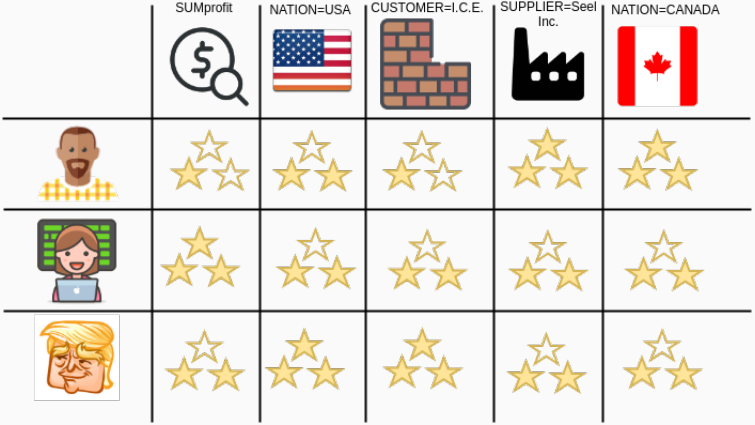
University of Tours

What is Alice best next move?



In fact, it depends!

A very subjective question?



We would need to “brain dump” analysts

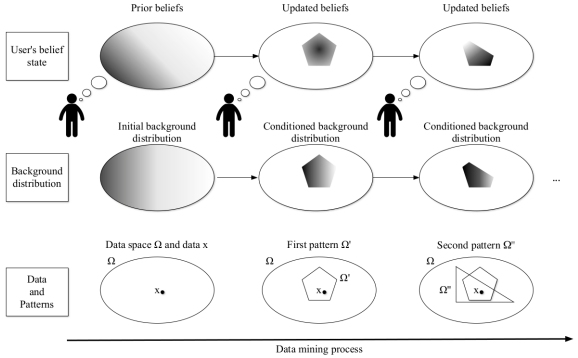
What is subjective interestingness?

- Objective interestingness
  - **user agnostic, based only on data**
  - generality, reliability, peculiarity, diversity and conciseness,
  - directly measurable evaluation metrics: support confidence, lift or chi-squared measures in the case of association rules
  - summaries: *compact descriptions of raw data at different concept levels (Geng & Hamilton)*
- Subjective interestingness
  - characterize the patterns' surprise and novelty when compared to **previous user knowledge** or **expected data distribution**
  - user adaptive exploration
  - subjective interestingness for explorative data mining

De Bie's framework

- a **pattern p** $\approx$ restriction of data space
- a **belief(p)** $\approx$ prior knowledge as a **probability distribution** over the pattern space
- surprise(p) = $-log(belief(p))$

$$Interestingness(p) = \frac{surprise(p)}{|p|}$$



How to translate subjective interestingness to BI?

Two main problems:

- Define the ”pattern”
  - Cell?
  - Query?
  - Query parts?
- Learn the belief function
  - how to take into account the specificities of BI?
  - how can we decide that two pieces of information are related in BI?
    - do we consider the **usage** (the query logs)?
    - do we consider the **structure** (the DB schema)?

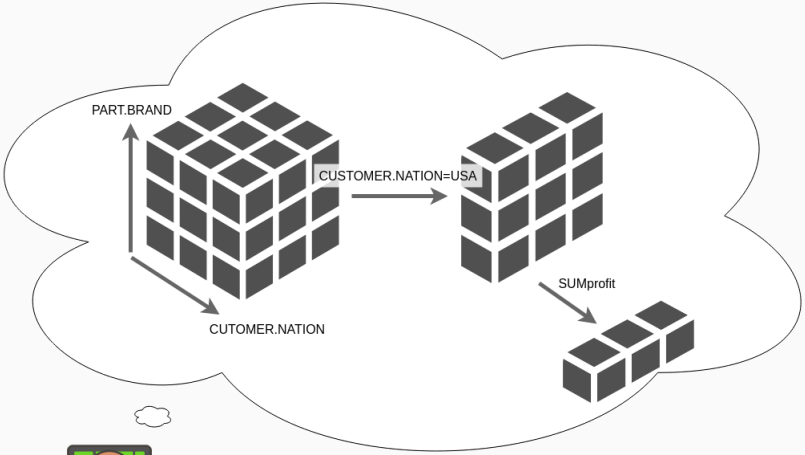
Our proposal

Belief expressed over query parts

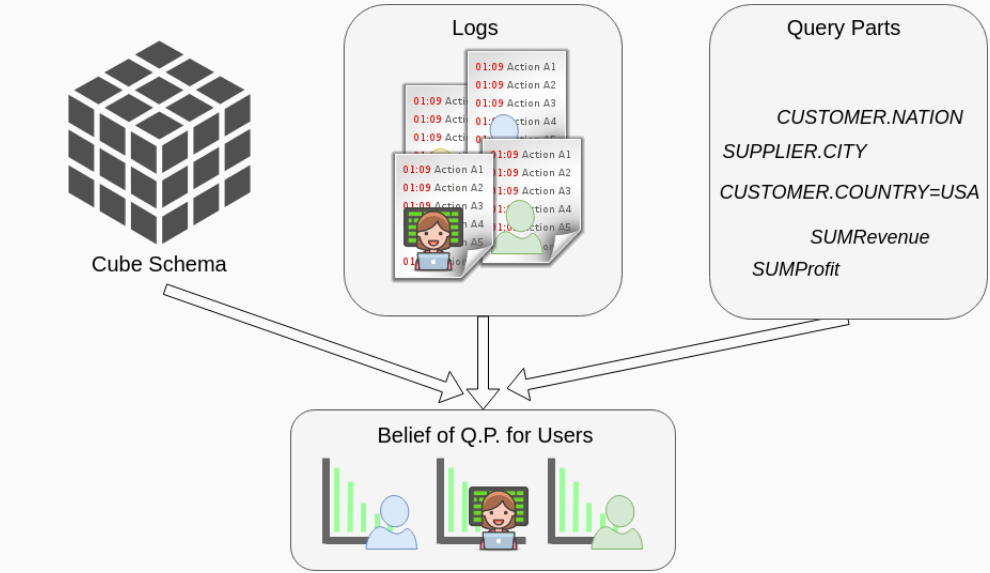
Classically, a query part is either:

- A group by set attribute
- A measure
- A selection predicate

Query parts as patterns



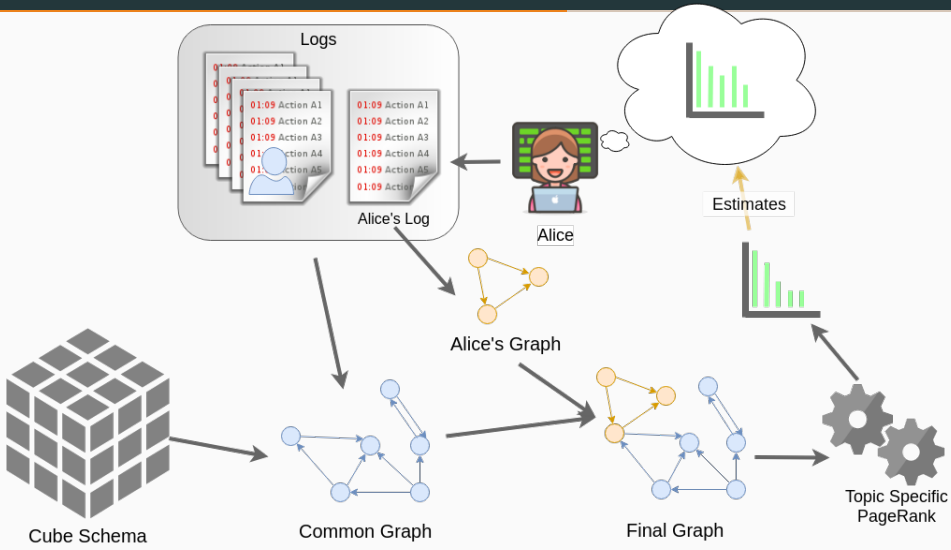
Our recipe so far



Random walk for learning the distribution

- consider a graph where vertices are query parts and edges are relations (precedence, co-occurrence) between them
- the user does a random walk over this graph
- the long term distribution of the user gives a measure of importance of the query parts
- it can be computed with a Page Rank
- or better, by a Topic-Specific Page Rank: a Page Rank where the user's query parts are more important than the others

Baking the pie



Experiments

Our ”Users”

- Artificial data generated with CubeLoad [1]
- mimic prototypical explorations
- More ”consistent” than real users
- Less noisy
- Only 4 profiles

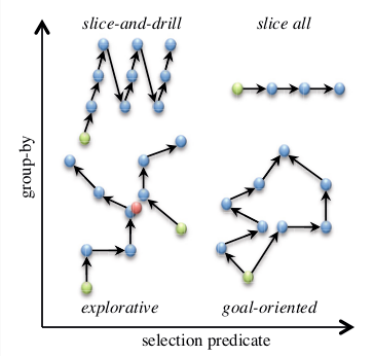
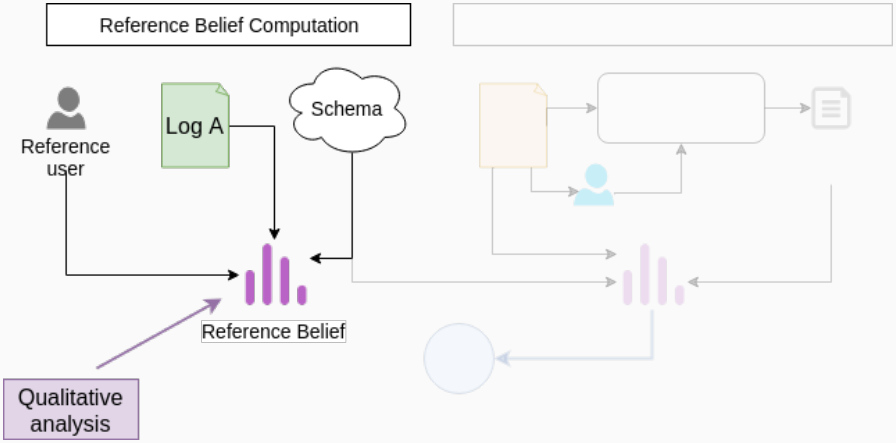


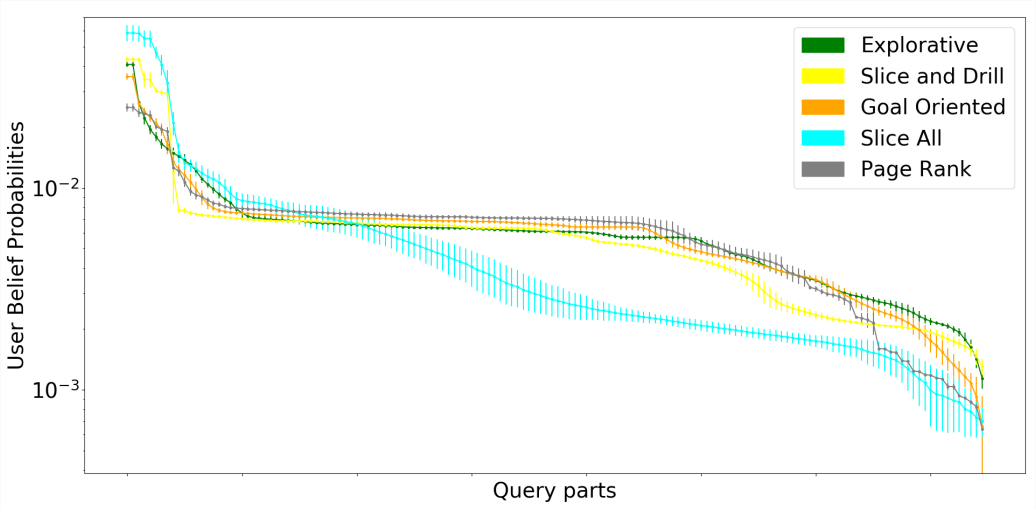
Figure 3: CubeLoad Templates

Protocol of the qualitative experiment

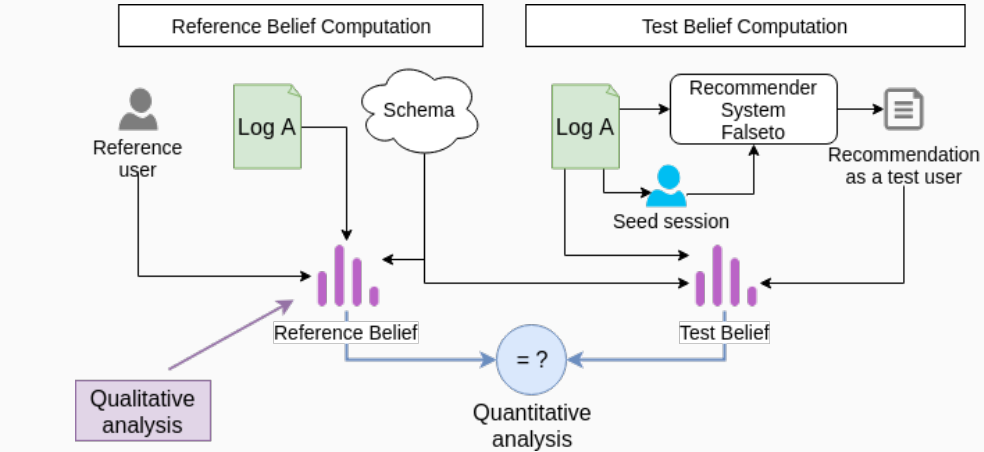


- determine if there is a belief profile that is representative of each CubeLoad template

Different user different beliefs



Protocol of the quantitative experiment



Introducing a user agnostic recommender in the loop

Robustness to logs exploring different regions (of the cube)

Observing a cognitive bubble

Tests\References	*Explorative*	*Goal Oriented*	*Slice All*	*Slice and Drill*
Explorative	**0.64**	0.60	0.47	0.60
Goal Oriented	0.64	**0.61**	0.46	0.60
Slice All	0.67	0.63	**0.47**	0.62
Slice and Drill	0.63	0.60	0.43	**0.58**

Explorative behavior

Conservative behavior
Cognitive bubble ?

Average Hellinger distance values on 10 runs when log files are identical

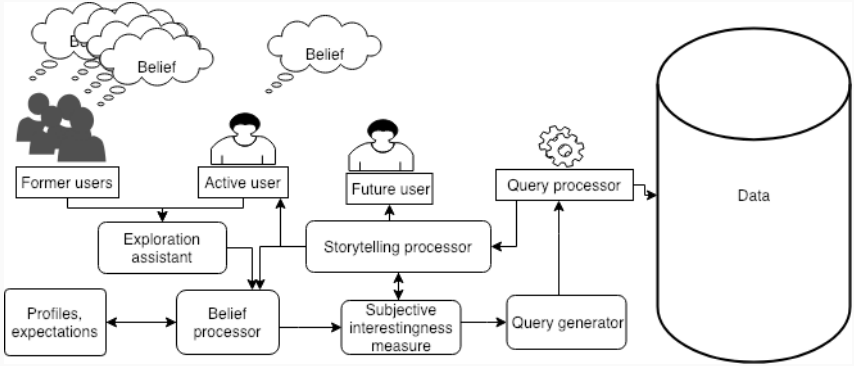
Conclusions

- First attempt to model belief in BI
- Capture potential relations between user knowledge as a graph
- $\Rightarrow$ use well-known Page-Rank for estimating probabilities
- Experiments
  - Different simulated user templates == different beliefs distributions
  - Possible detection of the cognitive bubble phenomena

On-going and Future work

- What about belief distribution over cell contents?
  - theoretically appealing but computationally painful...
  - (but we're on it)
- What about belief evolution along the exploration?
- Subjective interestingness is a trade-off between surprise and complexity of description
  - how to measure complexity of description in BI?
- How to validate a user "brain dump"?
  - Perform a user study based on an improved query recommender system with interestingness

Long term vision



Questions ?


S. Rizzi and E. Gallinucci.
Cubeload: A parametric generator of realistic OLAP workloads.
In *Advanced Information Systems Engineering - 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings*, pages 610–624, 2014.