CrowdMiner: Mining association rules from the crowd

Yael Amsterdamer Yael Grossman Tova Milo Pierre Senellart

TEL AVIV UNIVERSITY

TELECOM ParisTech

Introduction

- ***Crowd data sourcing*** collects data from the crowd, often by asking questions
- We want to learn about new domains from the crowd
  - E.g., health-related habits in some population
- Data is not recorded anywhere
- The contents of the domain are unknown
  - Discover what is **interesting** about this domain

What should we ask the crowd?

Data mining for the crowd?

- The discovery of data patterns in databases is done by **data mining.**
- Not suitable for our case
  - People do not remember enough details!

For example, it is unrealistic to expect people to remember every activity they did in the past, everything they have eaten, etc.

- They are far more likely to remember **personally prominent patterns**

“I drink red wine about once a week”

The model

We learn *association rules* of the form *a,b* → *c,d*

- E.g. , *“heartburn”* → *“baking soda”, “lemon”*

The answers contain

- **Rule support** – frequency of a,b,c,d
- **Rule confidence** – frequency of c,d given a,b
- **Items** (for an open question)

- **Significant rules** – average user support and confidence exceed fixed thresholds
- Users treated as random samples

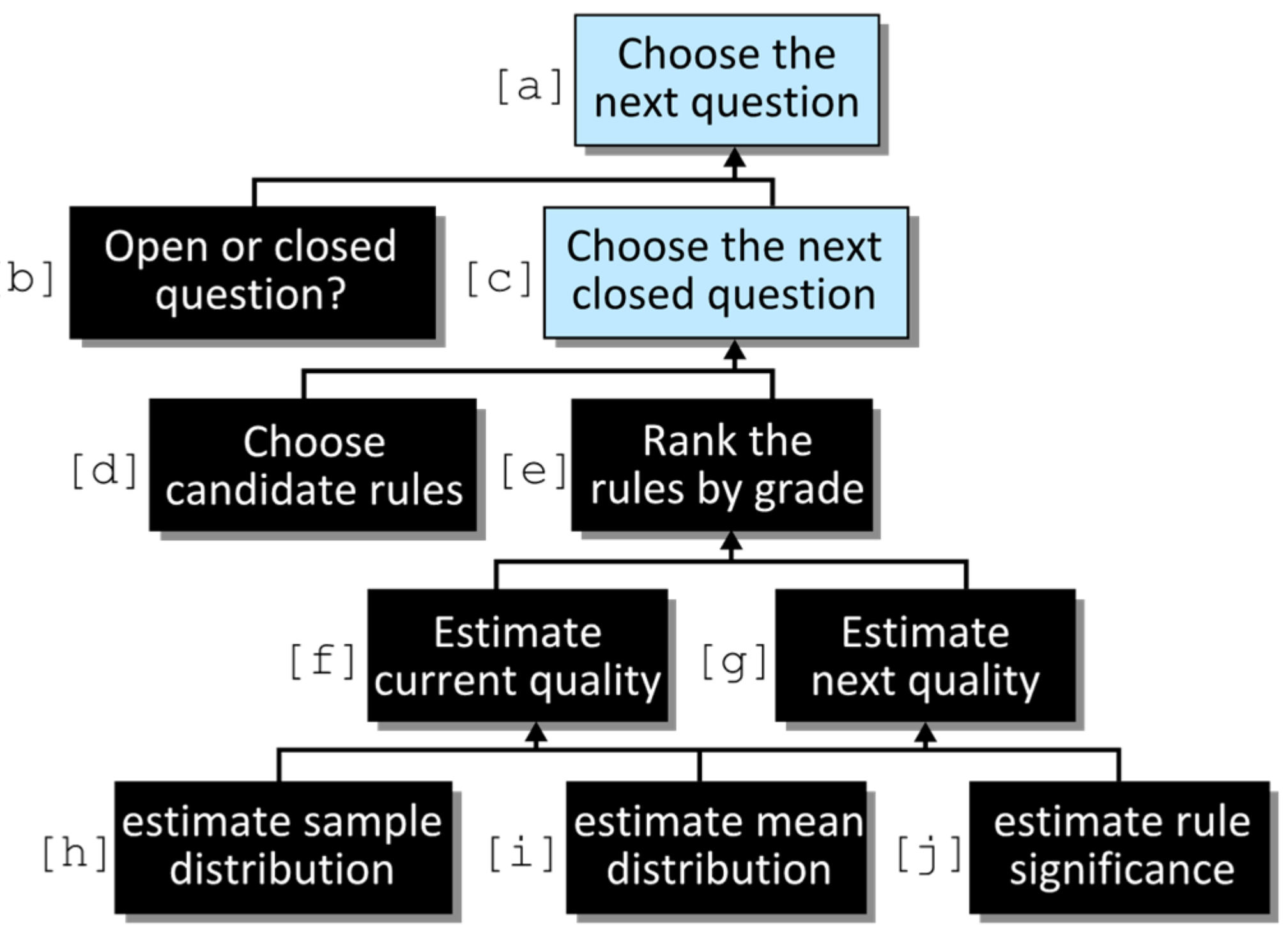
Our approach

- Use **personal summaries** to learn about **general trends**
- Treat individual answers as samples
- Combine two types of questions
  - ○ **Open questions** *“Complete: When I feel* tired *,*
    I usually go for a walk *”*
  - ○ **Closed questions** *“When you have a heartburn, do you take baking soda and lemon?”*
- Easier for users to answer
- Help digging deeper into their memories

We develop a system prototype *CrowdMiner* that interactively decides what to ask in order to discover significant data patterns

Choosing the Questions

A hierarchy of components that allow estimating the effect of the next question and choosing accordingly



[a] Choose the next question

[b] Open or closed question?

[c] Choose the next closed question

[d] Choose candidate rules

[e] Rank the rules by grade

[f] Estimate current quality

[g] Estimate next quality

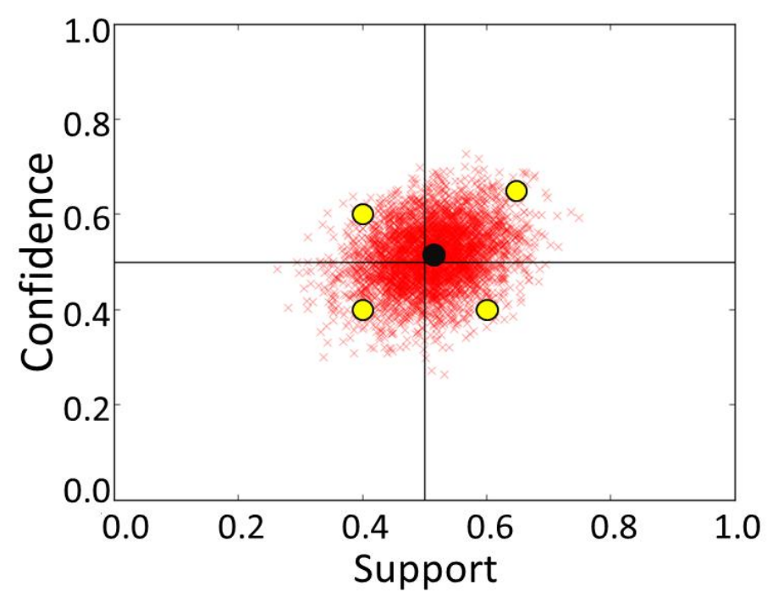
[h] estimate sample distribution

[i] estimate mean distribution

[j] estimate rule significance

Error Estimations

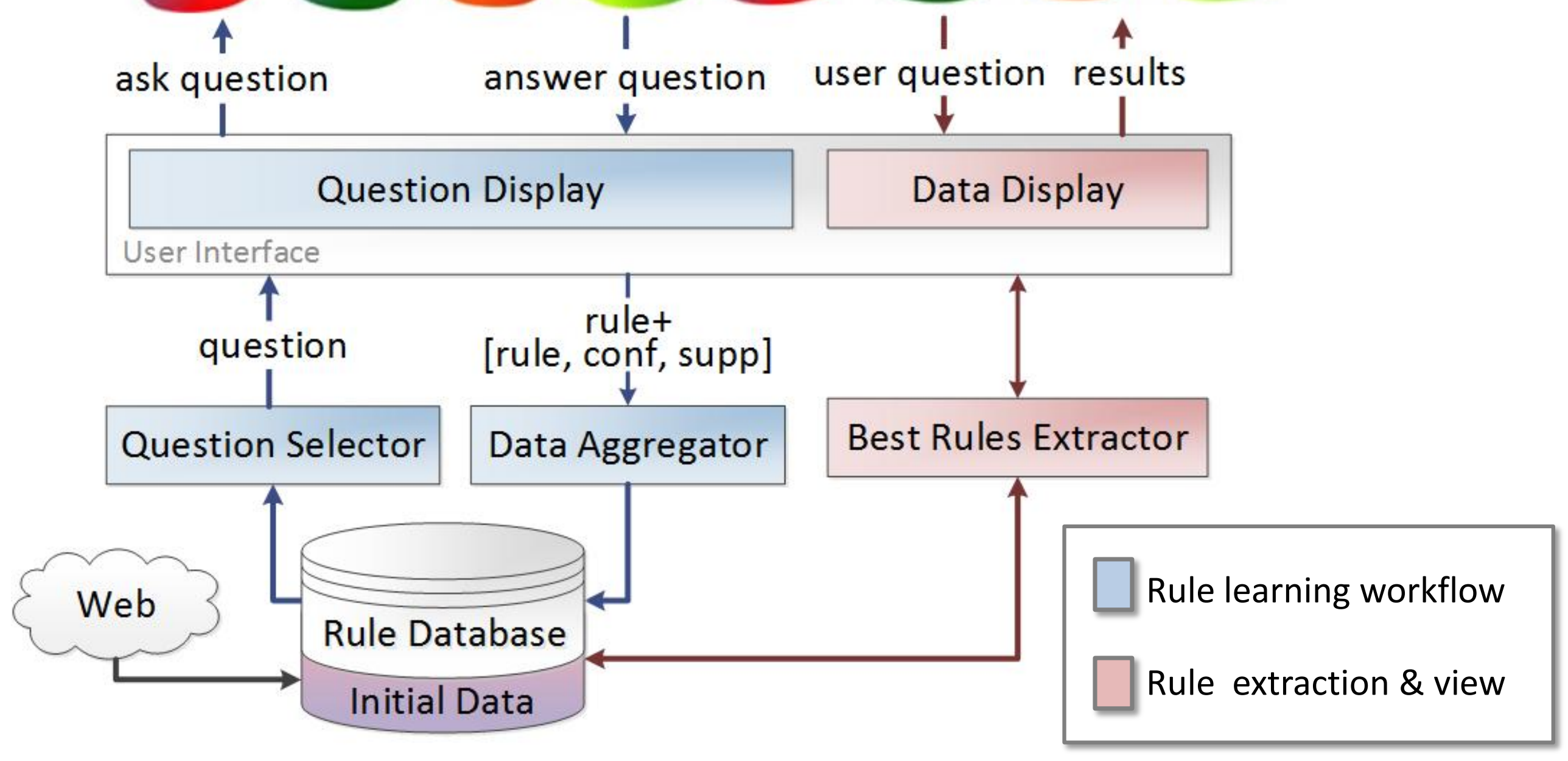
- Not all the users can be asked about every rule
- We want to estimate the probability of making an error – given the current knowledge
  - We learn a distribution of the answer support and confidence
  - **Significance estimation** – by the position of >0.5 of the distribution mass
  - **Error probability** – for the true mean to be on the other side of the thresholds
- The next question is the one expected to minimize the overall error



Well-Being Portal

- Learn about the **health habits** of others – by browsing the portal
  - Sports activities, eating habits, natural treatments
  - …
- Portal users are occasionally prompted with **questions**
  - About their personal habits
  - Computed by our algorithm
- User **answers** are processed to deduce rules (associations) between well-being concepts in the portal
- The portal allows browsing the learned rules

System Architecture


ask question
answer question
user question
results
Question Display
Data Display
User Interface
question
rule+
[rule, conf, supp]
Question Selector
Data Aggregator
Best Rules Extractor
Web
Rule Database
Initial Data
Rule learning workflow
Rule extraction & view