Crowd Mining

TEL AVIV UNIVERSITY

Yael Amsterdamer Yael Grossman Tova Milo Pierre Senellart

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., traditional (folk) medicine in some region
  - Or the leisure habits of hi-tech workers
- Data is not recorded anywhere
- The contents of the domain are unknown
  - Discover what is **interesting** in 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 folk healers to remember comprehensive details of all the cases they have treated in the past.

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

"I treat patients with a cold every week"

The model

We learn *association rules* of the form $a,b \rightarrow c,d$

- E.g. , *"heartburn"* $\rightarrow$ *"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 are sampled uniformly at random

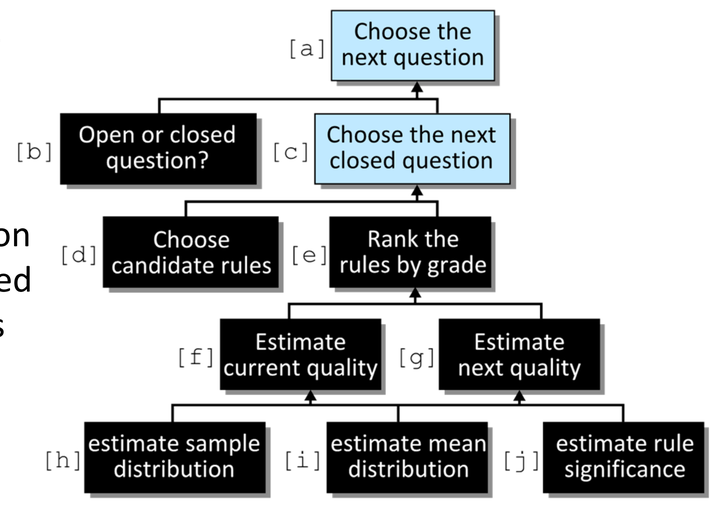
Our approach

- Use **personal summaries** to learn about **general trends**
- Treat individual answers as samples
- Combine two types of questions
  - **Open questions** — *"Which symptoms do you usually encounter?''*
  - **Closed questions** — *"Do you use baking soda and lemon to relieve a heartburn?"*
- Easier for users to answer
- Help digging deeper into their memories

We propose a **formal model**, a generic crowd mining **framework**, effective **implementation** for framework components and an **experimental study**

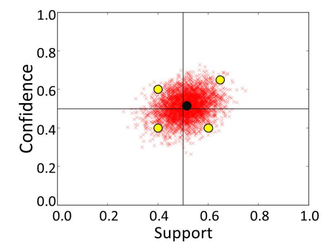
Component framework

A hierarchy of black-box components whose implementation may be adapted to the settings



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



Experiments

- 3 **new benchmark datasets** (with known ground truth):
  - Synthetic
  - Retail (market basket analysis)
  - Wikipedia editing records
- A system *CrowdMiner* and two baseline alternatives
  - Random
  - Greedy
- Varying the parameters, such as the mixture of open and closed questions, prior knowledge etc.