Crowd Mining

Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart

TEL AVIV UNIVERSITY

TELECOM ParisTech

Crowd data sourcing - Background

- Outsourcing data collection to the crowd of Web users
  - When people can provide the data
  - When people are the only source of data
  - When people can efficiently clean and/or organize the data

Crowdsourcing in an open world

- Human knowledge forms an open world
- Assume we know nothing, e.g., on folk medicine
- We would like to find what is *interesting* and *important* about folk medicine practices around the world.

What questions should be asked?

Back to classic settings

- Significant data patterns are identified using data mining techniques.
- Consider: *association rules*
  - E.g., *"heartburn"* → *"lemon", "baking soda"*
- Queries are dynamically constructed in the course of the learning process
- **Is it possible to mine the crowd?**

Asking the crowd

Let us model the history of every user as a *personal database*

Treated a sore throat with garlic and oregano leaves…
Treated a sore throat and low fever with garlic and ginger …
Treated a heartburn with water, baking soda and lemon…
Treated nausea with ginger, the patient experienced sleepiness…
…

- Every case = a *transaction* consisting of *items*
- Not recorded anywhere – a hidden DB
- It is hard for people to recall many details about many transactions!

But,
they can often provide summaries, in the form of personal rules

- *To treat a sore throat I often use garlic*
- *Interpretation: "sore throat" → "garlic"*

Two types of questions

- Free recollection (mostly simple, prominent patterns)

→ **Open questions**

Tell me how you treat a particular illness

"I typically treat <u>nausea</u> with <u>ginger infusion</u>"

- Targeted questions (may be more complex)

→ **Closed questions**

When a patient has both <u>headaches</u> and <u>fever</u>, how often do you use a <u>willow tree bark</u> infusion?

We use the two types interleavingly.

Personal Rules

- If people know which rules apply to them, why mine them?
  - Personal rules may or may not indicate general trends
  - Concrete questions help digging deeper into users' memory

Crowd Mining - Contributions (at a very high level)

- **Formal model** for crowd mining.
- **A Framework** of the generic components required for mining the crowd
- **Significance and error estimations.** Given the knowledge collected from the crowd, which rules are likely to be significant and what is the probability that we are wrong.
  [and, how will this change if we ask more questions…]
- **Crowd-mining algorithm**. Iteratively choosing the best crowd question and estimating significance and error.
- **Implementation & benchmark.**

The model: User support and confidence

- A set of users U
- Each user $u \in U$ has a (hidden) transaction database D_u
- Each rule $X \rightarrow Y$ is associated with:

user support

$$\mathrm{supp}_u(X \to Y) := \frac{|\{t \in D_u | X \cup Y \subseteq t\}|}{|D_u|}$$

user confidence

$$\mathrm{conf}_u(X \to Y) := \frac{|\{t \in D_u | X \cup Y \subseteq t\}|}{|\{t \in D_u | X \subseteq t\}|}$$

Model for closed and open questions

- **Closed questions:** X →$^{?}$ Y
  - **Answer:** (approximate) user support and confidence
- **Open questions:** ? →$^{?}$?
  - **Answer:** an arbitrary rule with its user support and confidence

"I typically have a headache once a week. In 90% of the times, coffee helps.

$$\text{supp}_u(\text{headache} \rightarrow \text{coffee}) = \frac{1}{7} \cdot \frac{9}{10} \qquad \text{conf}_u(\text{headache} \rightarrow \text{coffee}) = \frac{9}{10}$$

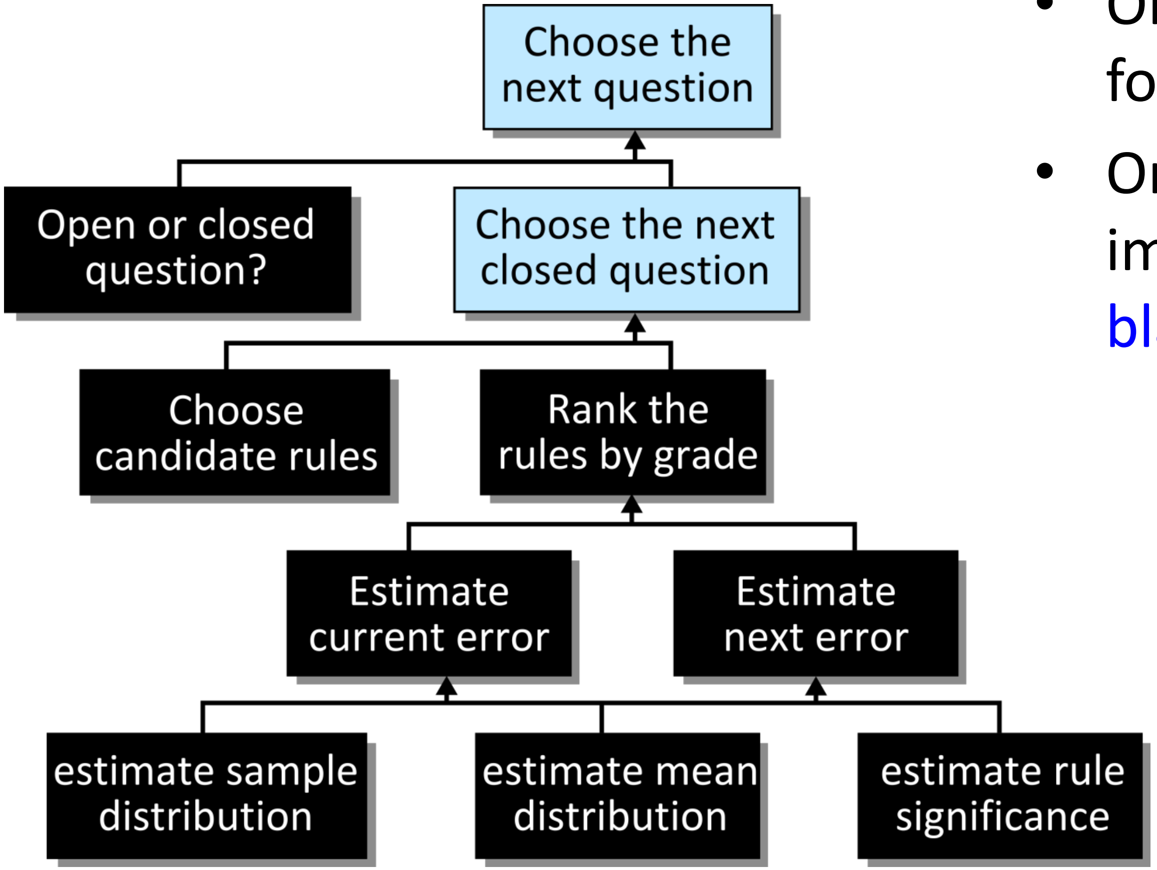
Significant rules

- Overall support and confidence defined as the mean user support and confidence

- Significant rules are those whose overall support and confidence are both above specified thresholds Θ_s, Θ_c.

- Goal: estimating rule significance while asking the smallest possible number of questions to the crowd

Framework components



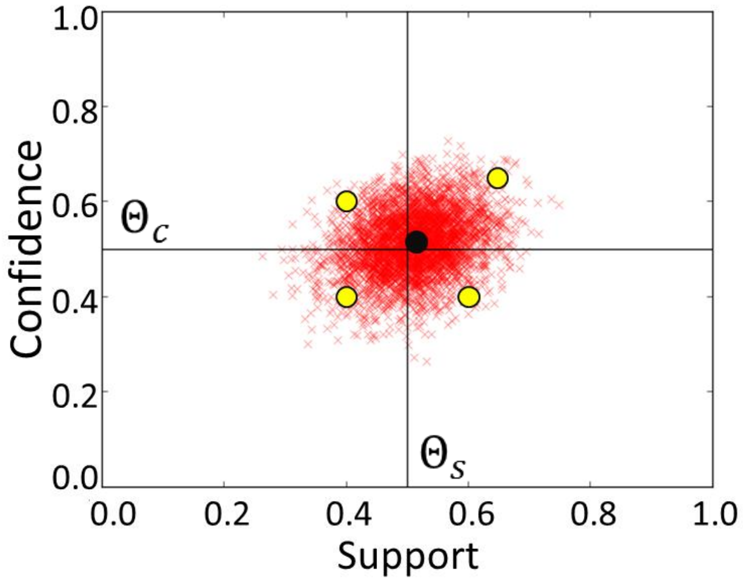
- One generic framework for crowd-mining
- One particular choice of implementation of all black boxes

Estimating the mean distribution

- Treating the current answers as a random sample of a hidden distribution g_r, we can approximate the distribution of the hidden mean f_r
- μ – the sample average
- Σ – the sample covariance
- K – the number of collected samples

$$f_r \sim N\left(\mu, \frac{\Sigma}{K}\right)$$

- In a similar manner we estimate the hidden distribution g_r



Rule Significance and error probability

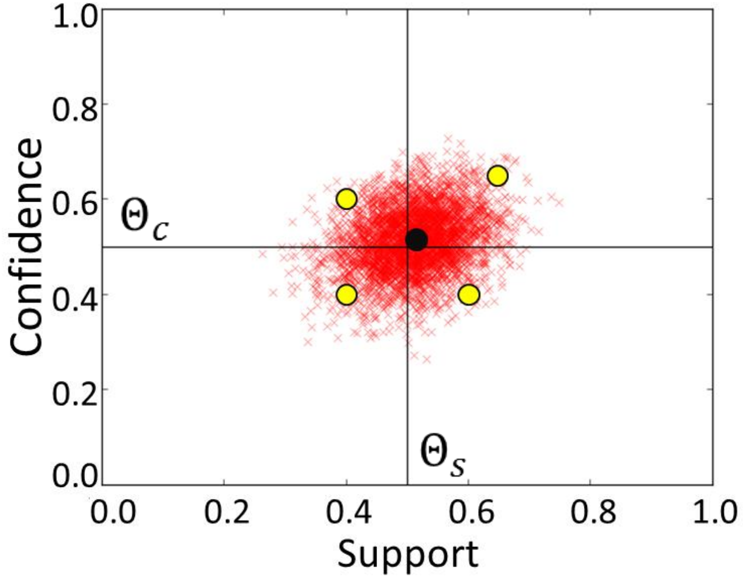
- Define M_r as the probability mass above both thresholds for rule r

$$M_r = \int_{\Theta_s}^{1} \int_{\Theta_c}^{1} f_r(s,c)\mathrm{d}c\mathrm{d}s$$

- r is significant iff M_r is greater than 0.5

- The error probability is

$$P_{\mathrm{err}}(r) = \min\{M_r, 1 - M_r\}$$

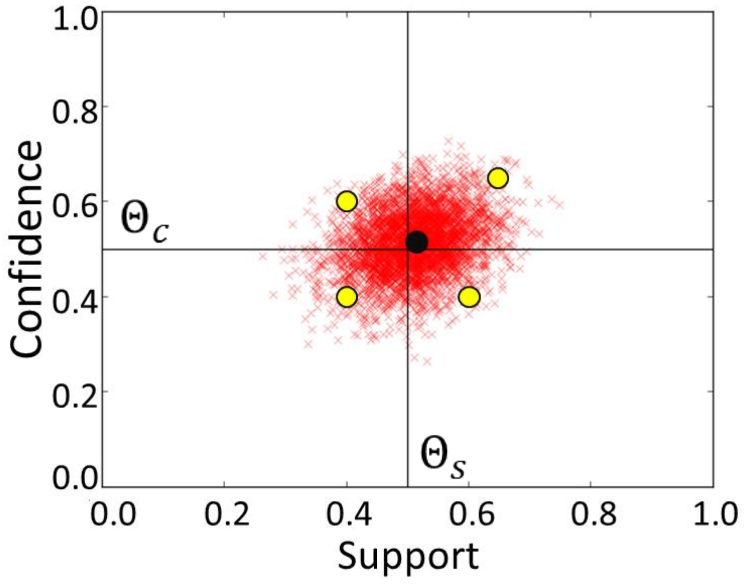


The next error probability

- The current distribution g_r for some rule can also be used for estimating what the next answer would be
- We integrate the resulting error probability over the possible next answers, to get the expected next error $\mathrm{E}[P'_{err}(r)]$

- **Optimization problem:** The best rule to ask about leads to the best output quality
- For quality := *overall error*, this is the rule that induces the largest error reduction

$$\operatorname{argmax}_{r \in R} P_{err}(r) - \mathrm{E}[P'_{err}(r)]$$



Completing the picture

- Which rules should be considered as candidates for the next question?
  - Small rules, rules similar to significant rules are most likely to be significant
  - Similarly to classic data mining

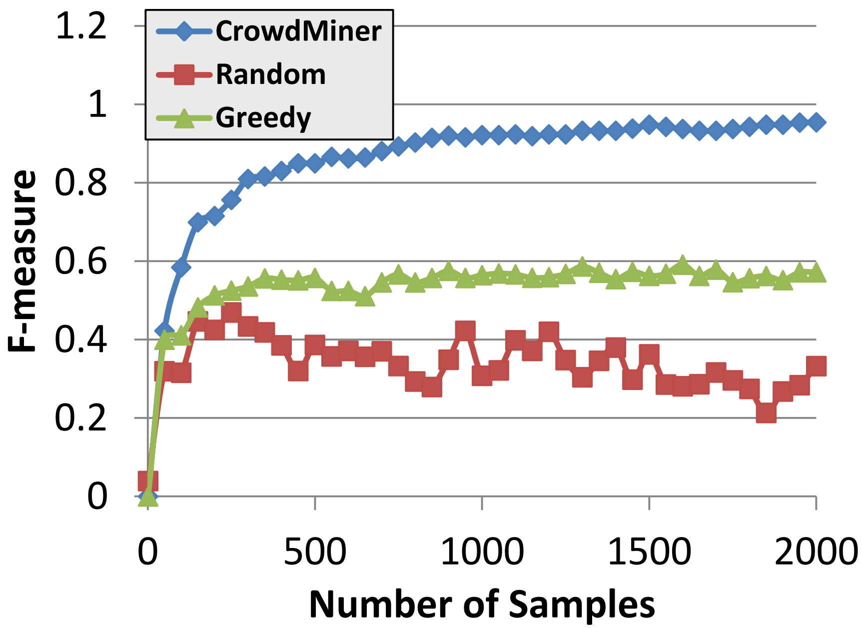
- Should we ask an open or closed question?
  - Keeping a fixed ratio of open questions balances the tradeoff between precision and recall
  - Similarly to sequential sampling

Experiments

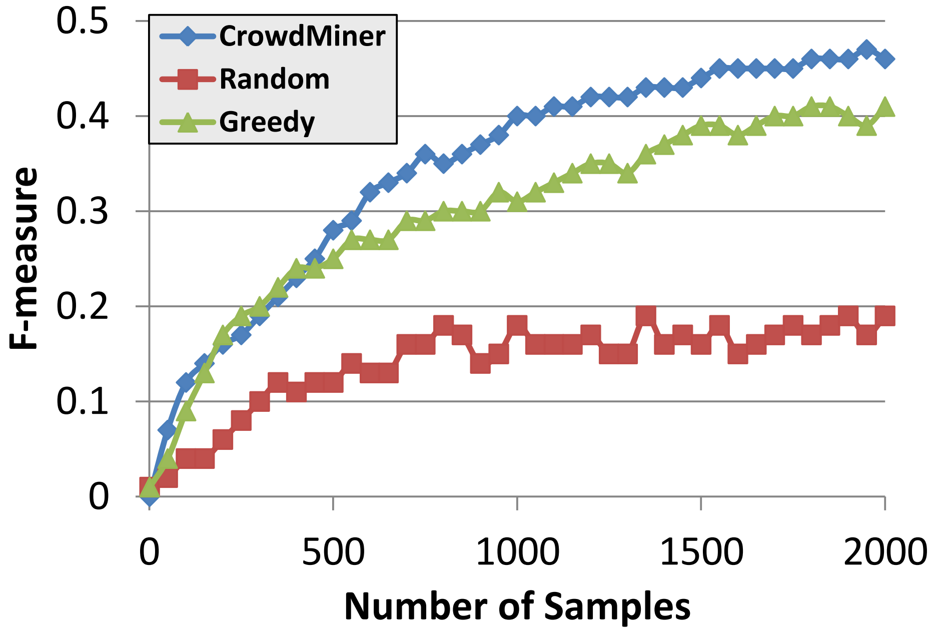
- 3 new benchmark datasets
  - Synthetic
  - Retail (market basket analysis)
  - Wikipedia editing records

- A system prototype, *CrowdMiner, and* 2 baseline alternatives
  - Random
  - Greedy (that asks about the rules with fewest answers)

Experimental Results

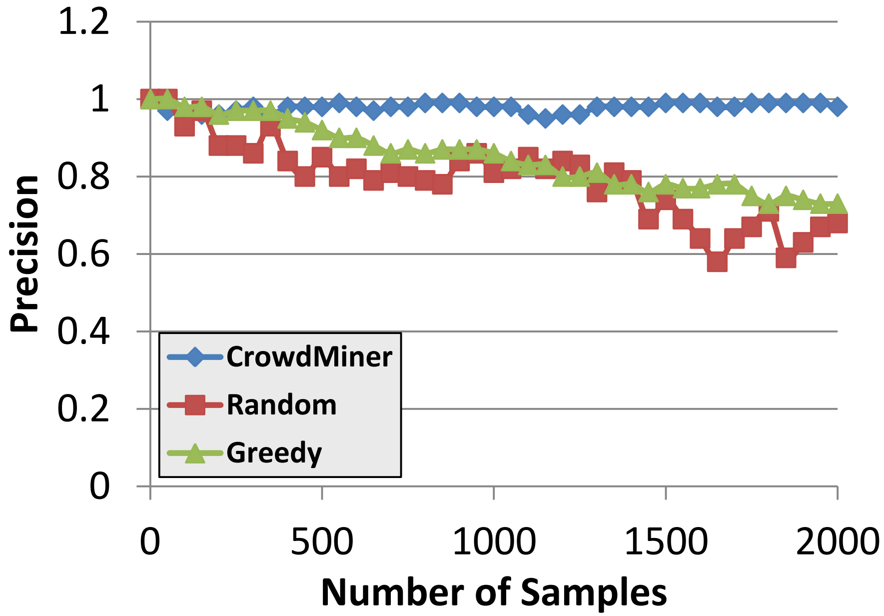


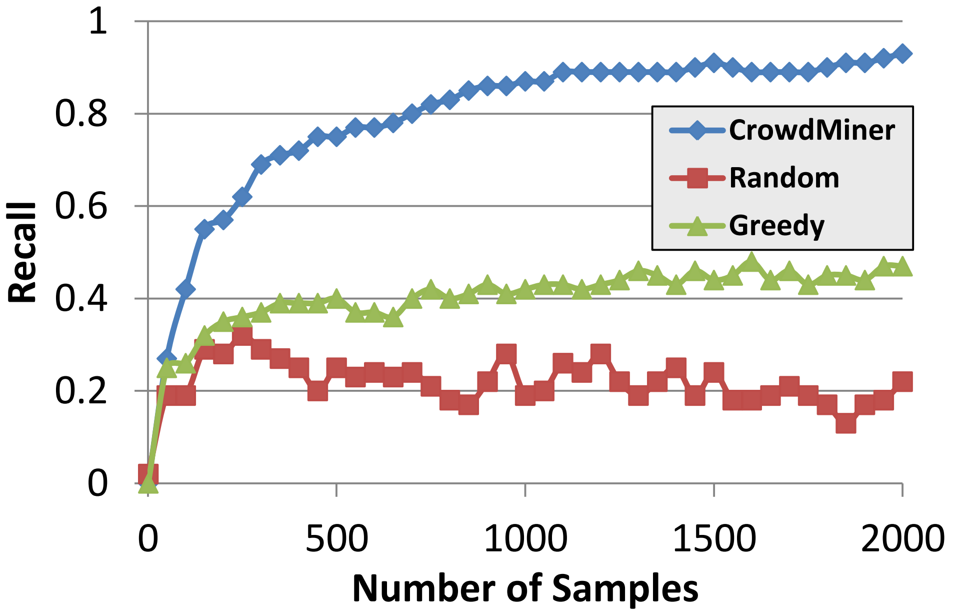
Retail Dataset



Wikipedia Dataset

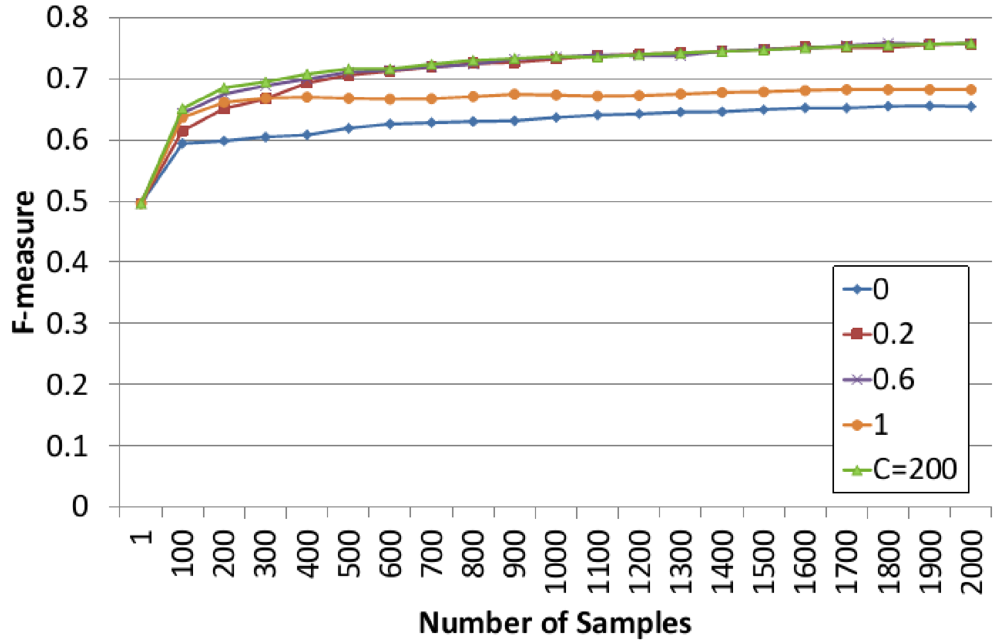
Experimental Results





- Better precision – Greedy loses precision as new rules are explored
- Much better recall – due to adding new rules as candidates.

Experimental Results



- An open questions ratio of 0.2-0.6 yields the best quality

Summary

- The goal: learning about new domains from the crowd
- By identifying significant data patterns
- Data mining techniques cannot be used as-is
- Our solution includes
  - A **model** for the crowd behavior
  - A crowd mining **framework** and concrete component **implementations**
  - Benchmark datasets and a prototype system *CrowdMiner* used for experimentation

Related work

- **Declarative crowdsourcing frameworks** [e.g., Doan et. Al PVLDB'11, Franklin et. Al SIGMOD'11, Marcus et. Al CIDR'11, Parameswaran et. Al CIDR'11]
  - We consider identifying patterns in unknown domains
- **Association rule learning** [e.g., Agrawal et. Al VLDB'94, Toivonen VLDB'96, Zaki et. Al RIDE'97]
  - Transactions are not available in our context, sampling rules does not perform as well as interleaving closed and open questions
- **Active Learning** [e.g., Dekel et. Al COLT'09, Sheng et. Al SIGKDD'08, Yan et. Al ICML'11]
  - In our context every user has a partial picture, no "right" or ``wrong''
- **Sequential Sampling** [Vermorel et. Al ECML'05]
  - Combining the exploration of new knowledge with the exploitation of collected knowledge

Ongoing and Future work

- **Leveraging on rule dependencies**
  - From an answer on one rule we can learn about many others
  - Semantic dependencies between rules
- **Leveraging on user info**
- **Other types of data patterns**
  - Sequences, action charts, complex relationships between items
- **Mining given a query**
  - Data mining query languages
- … and many more

Thank You!

Questions?