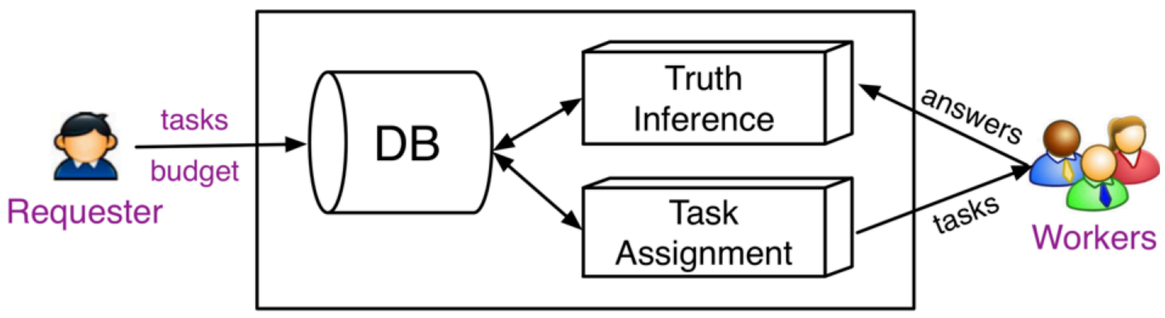
DOCS: A Domain-Aware Crowdsourcing System Using Knowledge Bases

Yudian Zheng*, Guoliang Li#, Reynold Cheng*

*University of Hong Kong, #Tsinghua University

Crowdsourcing Workflow

- **Task Assignment:** When a worker comes to the platform, the worker will be assigned to a set of tasks;
- **Truth Inference:** When a worker accomplishes tasks, the platform will collect answers from the worker.

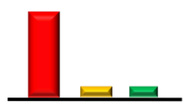


Existing Works Fail in QA Tasks

- Each task is related to different domains

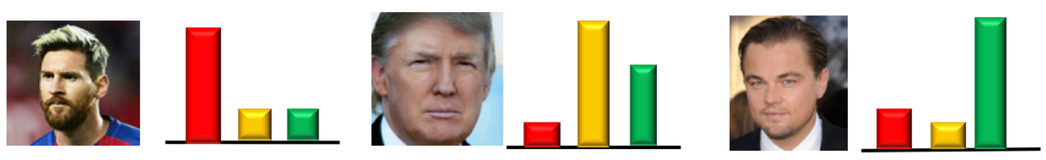
Sports Politics Entertainment

Did Michael Jordan win more NBA championships than Kobe Bryant? → Sports



Is there a name for the song that FC Barcelona is known for? → Sports & Entertainment

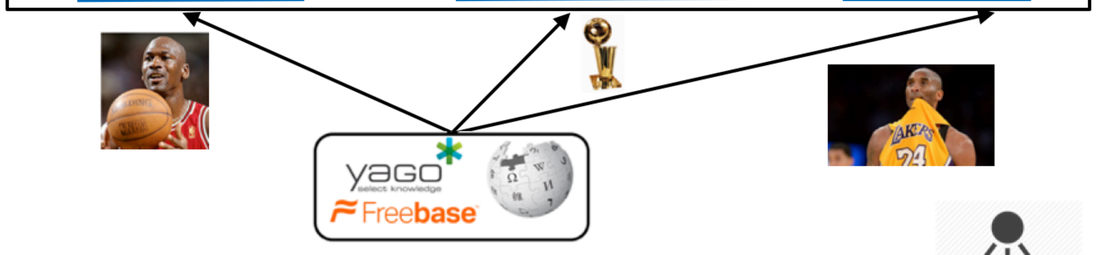
- Each worker has diverse qualities over domains



Build Domain Aware Task Model

- (1) Entity linking (map entity to knowledge bases)

Did Michael Jordan win more NBA championships than Kobe Bryant?

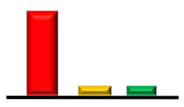


- (2) Hierarchical domains in knowledge bases

- (3) Obtain the task model (a vector of distribution)

Sports Politics Entertainment

Did Michael Jordan win more NBA championships than Kobe Bryant? → Sports



Build Domain Aware Worker Model

- Use qualification test (like an "exam") amazon mechanical turk Artificial Artificial Intelligence
- Two rules for selecting qualification test
  (1) Each selected task should capture a certain domain
  (2) The domain distribution of selected tasks should approximate the distribution of all tasks

KL-divergence

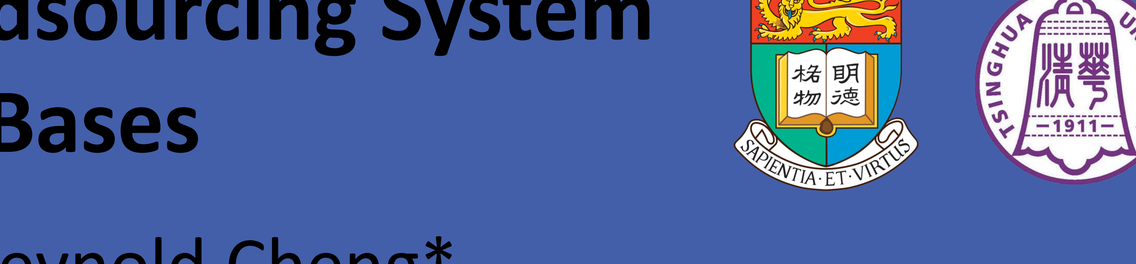
$$\min_{\{n'_k\}} \sum_{k=1}^{m} \frac{n'_k}{n'} \cdot \ln \frac{n'_k \cdot n}{n' \cdot \sum_{i=1}^{n} r_k^{t_i}}$$

$$\text{s.t. } \sum_{k=1}^{m} n'_k = n' \text{ and } n'_k \in \mathbb{N} \text{ for } 1 \le k \le m.$$

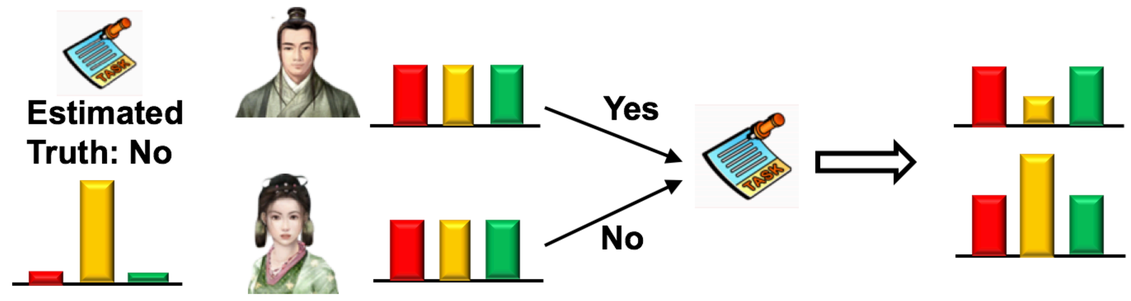
Truth Inference

- 1. Quality for each worker ⇒ Truth for each task

Sports Politics Entertainment



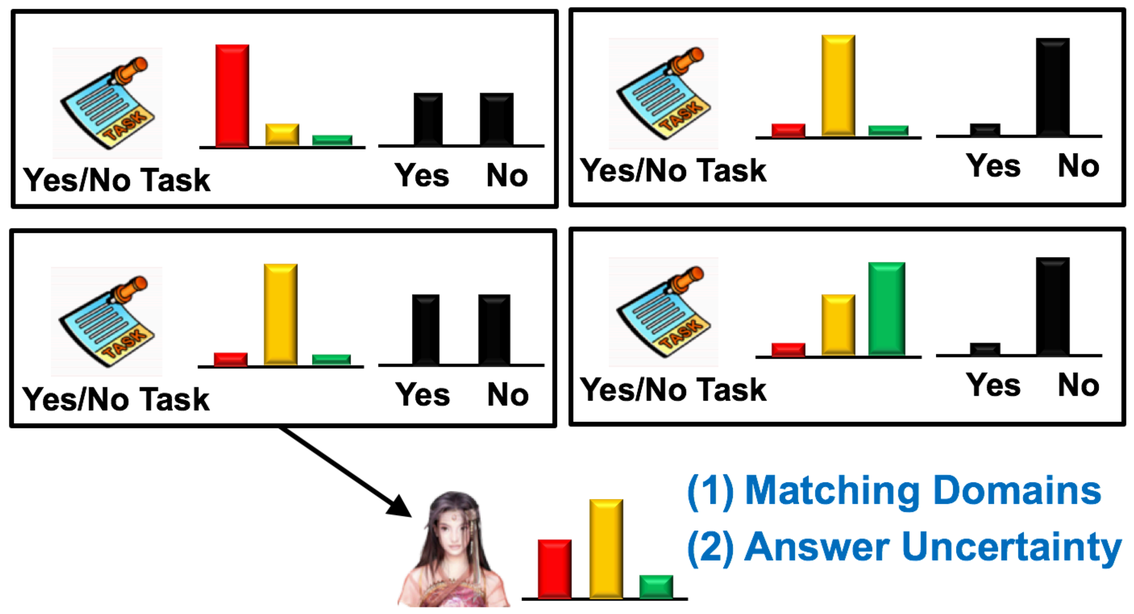
- 2. Truth for each task ⇒ Quality for each worker



Task Assignment

- Select the most suitable tasks for assignment

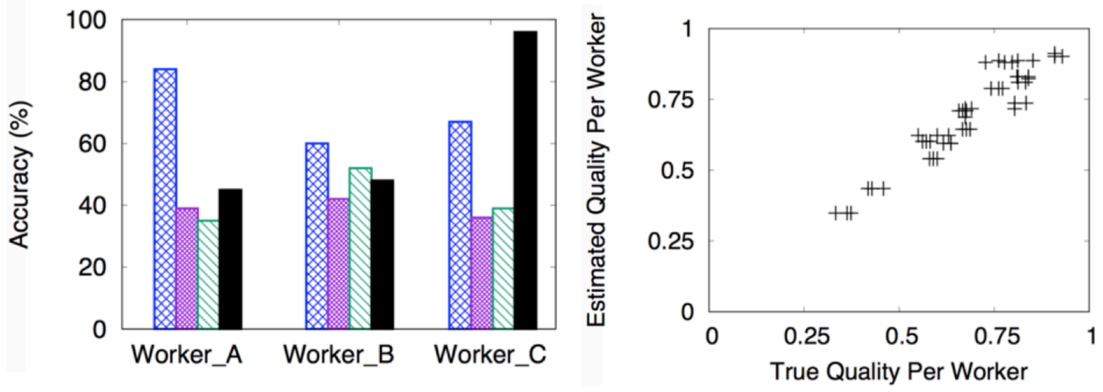
Sports Politics Entertainment



(1) Matching Domains
(2) Answer Uncertainty

Experiments

- Worker Characteristics on Dataset D_Item



- System Comparisons