DOCS: A Domain-Aware Crowdsourcing System Using Knowledge Bases

Yudian Zheng, Guoliang Li, Reynold Cheng

University of Hong Kong, Tsinghua University
Crowdsourcing Workflow

- **Requester** deploys tasks and budget on crowdsourcing platform (e.g., AMT)
- **Workers** interact with platform (2 phases)
  - **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.
Did Michael Jordan win more NBA championships than Kobe Bryant?

Is there a name for the song that FC Barcelona is known for?
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?

Is there a name for the song that FC Barcelona is known for?

- Each worker has diverse qualities over domains
Did Michael Jordan win more NBA championships than Kobe Bryant?
Part I: Domain Aware Task Model (3 steps)

○ Step 1: Entity linking (map entity to knowledge bases)

Did Michael Jordan win more NBA championships than Kobe Bryant?

Part I: Domain Aware Task Model (3 steps)

- **Step 2**: Hierarchical domains in knowledge bases

- **Step 3**: For each task, we obtain the task model (a vector of distribution)

Did Michael Jordan win more NBA championships than Kobe Bryant?
Part II: Domain Aware Worker Model

- Model each worker using a vector
  - Sports
  - Politics
  - Entertainment

Each element in the vector is in the range of \((0,1)\), indicating the expertise of the worker to a specific domain.

- Our ideas to initialize each worker’s model

  Use qualification test (like an “exam”) i.e., assign the tasks (with known truth) to the worker when the worker comes at first time.
Two rules for selecting qualification test

(1) Each selected task should capture a certain domain

Did Michael Jordan win more NBA championships than Kobe Bryant? Good: only related to one domain (sports)

Is there a name for the song that FC Barcelona is known for? Bad: related to multiple domains (both sports & entertainment)

(2) The domain distribution of selected tasks should approximate the distribution of all tasks

KL-divergence

\[
\min_{\{n'_k\}} \sum_{k=1}^{m} \frac{n'_k}{n'} \cdot \ln \frac{n'_k \cdot n}{n' \cdot \sum_{i=1}^{n} r_i^t}
\]

s.t. \( \sum_{k=1}^{m} n'_k = n' \) and \( n'_k \in \mathbb{N} \) for \( 1 \leq k \leq m \).
Truth Inference

1. **Quality for each worker** \(\rightarrow\) **Truth for each task**

- **Domain**
  - Yes/No Task
  - Sports
  - Politics
  - Entertainment

- **Quality**
  - Answer: Yes
  - No

- **Truth**
  - Yes
  - No
Truth Inference (cont’d)

2. Truth for each task $\Rightarrow$ Quality for each worker

- Sports
- Politics
- Entertainment

- Estimated Truth: No

- Quality

- Answer

- Yes

- No
Task Assignment

- Select the most suitable tasks for assignment

[Bar charts showing task assignment]

(1) Matching Domains
(2) Answer Uncertainty
# Experiments

## Dataset Setting (1)

<table>
<thead>
<tr>
<th>Name</th>
<th>#Tasks</th>
<th>Domains</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_Item: ItemCompare Dataset</td>
<td>360</td>
<td>NBA, Food, Auto, Country</td>
<td>It asks workers to compare between two items</td>
<td>Which food contains more calories, Chocolate or Honey?</td>
</tr>
<tr>
<td>D_4D: 4 Domain Dataset</td>
<td>400</td>
<td>NBA, Car, Film, Mountain</td>
<td>It asks workers about tasks on a certain domain</td>
<td>Did Michael Jordan win more NBA championships than Kobe Bryant?</td>
</tr>
</tbody>
</table>
Experiments

○ Dataset Setting (2)

<table>
<thead>
<tr>
<th>Name</th>
<th>#Tasks</th>
<th>Domains</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_QA: Yahoo QA Dataset</td>
<td>1000</td>
<td>Domains in Yahoo Answers</td>
<td>It asks workers tasks on Yahoo Answers</td>
<td>Where does chili originate from, Texas or Turkey?</td>
</tr>
<tr>
<td>D_SFV: SFV Dataset</td>
<td>328</td>
<td>Domains in Yahoo Answers</td>
<td>It asks workers the attribute of a person, where the answers are collected from different QA systems</td>
<td>What is the age of Bill Gates?</td>
</tr>
<tr>
<td>(a NLP dataset)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- Worker Characteristics on Dataset D_Item

Each worker has diverse quality over different domains. Estimated worker quality is close to real worker quality.
Experiments

- **System Comparisons**

![Bar chart comparing system performance across different datasets](chart.png)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>AskIt!</th>
<th>iCrowd</th>
<th>QASCA</th>
<th>Greedy</th>
<th>DOCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D_4D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D_QA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D_SFV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Summary

- Consider the domain aware task model and worker model

- Design solutions to accurately estimate the task model and worker model

- Incorporate task model and worker model in truth inference and task assignment
THANKS FOR WATCHING

Q&A

Yudian Zheng, Guoliang Li, Reynold Cheng

University of Hong Kong, Tsinghua University