Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge

Eddy Maddalena*, Marco Basaldella*, Dario De Nart*, Dante Degl’Innocenti*, Stefano Mizzaro*, Gianluca Demartini+

* University of Udine, Italy
+ University of Sheffield, UK
Crowdsourcing Relevance Judgements

• Task: Given a Query, Document pair

  Is the doc

  highly relevant, relevant, partially relevant, not relevant?

• Ask multiple workers

• Aggregate answers to obtain a relevance label

Query: jaguar
Our Research Question

Can we limit the time to judge to reduce the cost ($$) of creating IR test collections?

Hypothesis
Yes, but with quality loss
Our Experimental Setup

• **TREC8** Topics and documents (binary and 4-level expert judgements)
• **CrowdFlower**, repeated for USA and IND
• Majority vote aggregation
• Quality control: topic understanding question + high quality workers
• HIT Reward adapted based on the expected completion time

• Quality of a judgement: *Agreement* with editorial judgements
  • Cohen’s Kappa and distance with 4-level labels
Our Experimental Setup

• **E1 Unbound time** (i.e., the standard approach)
  • 5 judgements per doc, 8 documents, 5 topics, 2 crowds = 400 workers

• **E2** Document shown for a **predefined amount of time**
  • 30, 15, 7, 3 seconds. Each worker to judge 8 docs

• **E3 Same timeout** for all 8 documents (15 or 30 sec)

• **E4 Fixed budget**: comparison between
  • more quick judgements
  • few slow judgements
E1: We Have All the Time in the World

- RQ: **How much time** do crowd workers take to judge the relevance of a document **if no time constrain** is set?
  - 5 workers to judge a permutation of 8 docs

  - Median: 13 sec
  - Mean 24-25 sec

  a tail of very long execution times

  First doc takes longer (learning)
E1: We Have All the Time in the World

- No correlation of time with
  - Doc length
  - Doc readability
  - Topic
  - Relevance level
- Time vs Quality

![Graph showing time vs quality](image)

<table>
<thead>
<tr>
<th></th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>2.0</td>
<td>3.2</td>
<td><strong>5.1</strong></td>
<td>7.6</td>
<td>10</td>
<td>13</td>
<td>17</td>
<td>23</td>
<td>32</td>
<td>51</td>
<td>580</td>
</tr>
<tr>
<td>IN</td>
<td>1.9</td>
<td>3.4</td>
<td><strong>4.5</strong></td>
<td>7.0</td>
<td>9.9</td>
<td>13</td>
<td>17</td>
<td>22</td>
<td>31</td>
<td>46</td>
<td>630</td>
</tr>
</tbody>
</table>
E2: Faster! Faster! Sorry, Too Late

• Understand which is the **minimum amount of time required** to perform relevance judgments

• (max) timeouts: 30, 15, 7, 3 seconds

• Each worker to judge 8 docs, 2 for each timeout (one long, one short)

• Looking at Quality measures:
  • 3 and 7 secs are not enough
  • 15 slightly better than 30 (learning bias for position 1-2?)
E2: Faster! Faster! Sorry, Too Late

Time when document disappears

Time when judgement is made

Position of the document judged (1-8)

Variance across topics
E3: Selecting the Best Timeout

• We repeated E1 using 15 and 30 sec timeouts

• 15 seconds timeouts yield consistently better quality judgements
  • Than 30 seconds timeouts
  • Than no timeouts (E1 quality values)
Our Research Question

Can we limit the time to judge to reduce the cost ($$) of creating IR test collections?

Hypothesis

Yes, and it improves the quality!
Yes, but with quality loss
E4: Many Fast & Furious or a Few Laid-Back?

- **Fixed budget:**
  - small timeout, more workers
  - Long timeout, less workers

- We compared 10 combinations with the same budget

<table>
<thead>
<tr>
<th>Timeslot (sec)</th>
<th>6</th>
<th>7.9</th>
<th>10</th>
<th>13.7</th>
<th>16.7</th>
<th>21.5</th>
<th>25</th>
<th>30</th>
<th>37.5</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>25</td>
<td>19</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

- **Highest quality at 25-30 sec**
Findings

• The first couple of judgments done by a worker are of lower quality
• Judgements that take more than 30 seconds are of lower quality
• Time-outs in relevance judgements HITs can increase quality
• The best timeout to be used lies in the interval of 25-30 seconds and does not depend on topic, document, or crowd.
Conclusions

• Crowdsourcing Relevance Judgements for IR Evaluation can be expensive to scale

• Limiting the time to judge can control the cost

• But can also increase the quality!
  • By inducing workers to look at the document for a predefined amount of time
  • With a balance between boredom and stress -> “in the flow”

http://gianlucaademartini.net