The role of humans in crowdsourced semantics

Elena Simperl, University of Southampton*

WIC@WWW2014

*with contributions by Maribel Acosta, KIT

07 April 2014
Crowdsourcing Web semantics: the great challenge

- Crowdsourcing is increasingly used to augment the results of algorithms solving Semantic Web problems

- Research questions
  - Which form of crowdsourcing for what task?
  - How to design the crowdsourcing exercise?
  - How to combine different human- and machine-driven approaches?
There is crowdsourcing and crowsourcing...
Microtask crowdsourcing

Work is broken down into smaller (‘micro’) pieces that can be solved independently.
Hybrid systems (or 'social machines')

Physical World
(Network of social interactions)

Virtual world
(Participation and data supply)

Model of social interaction

Participation and data supply

Design and composition

Physical World
(Network of social interactions)

4/21/2014

Dave Robertson
Example: Hybrid data integration

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>email</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
<td>Mike</td>
<td>mike@a</td>
<td>ICDE-02</td>
</tr>
<tr>
<td>Social media</td>
<td>Jane</td>
<td>jane@b</td>
<td>PODS-05</td>
</tr>
</tbody>
</table>

Generate plausible matches
- paper = title, paper = author, paper = email, paper = venue
- conf = title, conf = author, conf = email, conf = venue

Ask users to verify
Does attribute paper match attribute author?

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>email</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
<td>Mike</td>
<td>mike@a</td>
</tr>
<tr>
<td>Social media</td>
<td>Jane</td>
<td>jane@b</td>
</tr>
</tbody>
</table>

McCann, Shen, Doan: Matching Schemas in Online Communities. ICDE, 2008
Example: Hybrid query processing

Use the crowd to answer DB-hard queries

Where to use the crowd:

- **Find missing data**
- **Make subjective comparisons**
- **Recognize patterns**

But not:

- Anything the computer already does well

Crowdsourcing Linked Data Quality Assessment
M Acosta, A Zaveri, E Simperl, D Kontokostas, S Auer, J Lehmann
The Semantic Web–ISWC 2013, 260-276

CROWDSOURCING LINKED DATA CURATION
Tasks to be crowdsourced

- **Incorrect object**
  - Example: `dbpedia:Dave_Dobbyn dbprop:dateOfBirth “3”`.

- **Incorrect data type or language tags**
  - Example: `dbpedia:Torishima_Izu_Islands foaf:name “鳥島”@en`.

- **Incorrect link to “external Web pages”**
Combination of approaches

Contest
LD Experts
*Difficult* task
Final prize

Microtasks
Workers
*Easy* task
Micropayments

Adapted from [Bernstein2010]
Workflow

1. Resource selection
   - [Manual]
   - [Per Class]
   - [Random]
   - Resource

2. Evaluation of resource’s triples
   - [No]
   - [Yes]
   - Selection of quality issues
   - List of incorrect triples classified by quality issue

3. HIT generation

4. Accept HIT
   - Assess triple according to the given quality issue
     - [Correct]
     - [Incorrect]
     - [Data doesn’t make sense]
     - [I don’t know]
     - [More triples to assess]

5. Submit HIT

(Find stage)
LD Experts in contest

(Verify stage)
Workers in paid microtasks
Microtask design

• Selection of foaf:name or rdfs:label to extract human-readable descriptions

• Values extracted automatically from Wikipedia infoboxes

• Link to the Wikipedia article via foaf:isPrimaryTopicOf

• Preview of external pages by implementing HTML iframe
Experiments

• **Crowdsourcing approaches:**
  - *Find* stage: Contest with LD experts
  - *Verify* stage: Microtasks (5 assignments)

• **Creation of a gold standard:**
  - Two of the authors of this paper (MA, AZ) generated the gold standard for all the triples obtained from the contest
  - Each author independently evaluated the triples
  - Conflicts were resolved via mutual agreement

• **Metric: precision**
  \[ p = \frac{TP}{TP + FP} \]
## Overall results

<table>
<thead>
<tr>
<th></th>
<th>LD Experts</th>
<th>Microtask workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of distinct participants</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>Total time</td>
<td>3 weeks (predefined)</td>
<td>4 days</td>
</tr>
<tr>
<td>Total triples evaluated</td>
<td>1,512</td>
<td>1,073</td>
</tr>
<tr>
<td>Total cost</td>
<td>~ US$ 400 (predefined)</td>
<td>~ US$ 43</td>
</tr>
</tbody>
</table>
Precision results: Incorrect object task

- MTurk workers can be used to reduce the error rates of LD experts for the Find stage

<table>
<thead>
<tr>
<th>Triples compared</th>
<th>LD Experts</th>
<th>MTurk (majority voting: n=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>509</td>
<td>0.7151</td>
<td>0.8977</td>
</tr>
</tbody>
</table>

- 117 DBpedia triples had **predicates related to dates** with incorrect/incomplete values:
  
  "2005 Six Nations Championship" Date 12.

- 52 DBpedia triples had **erroneous values from the source**:
  
  "English (programming language)" Influenced by ?.
Precision results: Incorrect data type task

<table>
<thead>
<tr>
<th>Data types</th>
<th>Experts TP</th>
<th>Experts FP</th>
<th>Crowd TP</th>
<th>Crowd FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>0.8270</td>
<td>0.4752</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number with decimals</td>
<td>341</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Millimetre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanometre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not specified / URI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• We analyzed the *189 misclassifications* by the *experts*:

![Pie chart showing the distribution of misclassifications: 50% Freebase links, 39% Wikipedia images, and 11% external links.]

• The *6% misclassifications* by the *workers* correspond to pages with a language different from English.
Summary of findings

• The effort of LD experts must be applied on those tasks demanding specific-domain skills.

• MTurk crowd was exceptionally good at performing data comparisons

• Lay users do not have the skills to solve domain-specific tasks, while experts performance is very low on tasks that demand an extra effort (e.g., checking an external page)