Crowdsourcing and Its Applications in Academic Researches

陳昇瑋

中央研究院 資訊科學研究所
Crowdsourcing

= Crowd + Outsourcing

“soliciting solutions via open calls to large-scale communities”
What tasks are crowdsourceable?
Software Development

Reward: 25,000 USD

Challenge Overview
The Seeker is measuring quantitative attribute data on experimental units arranged in rows and columns on a two-dimensional field, and is looking to improve and semi-automate a process of data quality review and data adjustment to acquire optimal, minimally biased estimates of the attributes of individuals. The Seeker currently has a manual process and is looking for the Solver to develop interactive software to facilitate the steps involved. More importantly, the Seeker would like the Solver to integrate an algorithm (or other tool) to identify for adjustment the surface patterns in the two-dimensional displays of measurement data that are not being addressed by the Seeker's current methods. Solvers with mathematical, engineering, physics, biology, or computer science backgrounds are encouraged to participate. The Seeker requests that Solvers document their recommended methodologies for solving this problem and submit demonstration code that implements their approach.
Data Entry

- Reward: 4.4 USD/hour
General Questions

- Reward: points on Yahoo! Answers

![Image of Yahoo! Answers question]

**Open Question**

*If you could live anywhere in the world, where would it be and why?*

Hi, I'm Samara Weaving and I have joined the cast of Home and Away in the role of Indigo.

Growing up, I had the opportunity to live in Singapore, Fiji and Indonesia as well as Australia. I particularly loved Singapore because it's so clean (gum is illegal there) and there is no winter.

I was wondering, if you could choose anywhere in the world to live, where would it be and why?

To learn more about me, check out my interview on the Home and Away website - http://au.tv.yahoo.com/home-and-away/features/article/-/article/5762754/samara-weaving/

1 month ago - 1 month left to answer.
Crowdsourcing in Academic Researches
Image Understanding

- Reward: 0.04 USD

main theme?
key objects?
unique attributes?
Crowdsourcing and Its Applications to Academic Researches / Sheng-Wei Chen

0.02 USD/task

find out photos of revolvers!
Image Understanding

Mechanical Turk Project
If you're using the turk, be sure to copy the text back into the HIT page so that you can be credited.
- Photo should be rotated 90 degrees left (counter-clockwise)
- Photo should be rotated 90 degrees right (clockwise)
- Photo should be turned upside down
- Photo is oriented properly
Please describe the picture in the box using 10 words or more:
абелла

Submit Turk
Skip / Load a different photo

The submit button MUST be clicked!

0.01 USD/ task
Human Action Recognition

0.01 USD/ task
0.01 USD/ task
Linguistic Annotations

- Word similarity (Snow et al. 2008)

Word pair similarity

Below is a list of pairs of words. For each pair, please assign a numerical similarity score between 0 and 10 (0 = words are totally unrelated, 10 = words are VERY closely related). By definition, the similarity of the word to itself should be 10. You may assign fractional scores (for example, 7.5).

boy    lad
Similarity (0-10)

coast  shore
Similarity (0-10)

USD 0.2 for labeling 30 word pairs
Linguistic Annotations

- Affection recognition (Snow et al. 2008)

  "Closing and cancellations top advice on flu outbreak"

USD 0.4 to label 20 headlines (140 labels)
Linguistic Annotations

- **Textual entailment**
  - If “Microsoft was established in Italy in 1985”, then “Was Microsoft established in 1985”? 

- **Word sense disambiguation**
  - “a bass on the line” vs. “a funky bass line”

- **Temporal annotation**
  - Ran happens before fell
More Examples

- **Document relevance evaluation**
  - Alonso et al. (2008)

- **Document rating collection**
  - Kittur et al. (2008)

- **Noun compound paraphrasing**
  - Nakov (2008)

- **Person name resolution**
  - Su et al. (2007)
Crowdsourcing: Good and Bad

- **Pros**
  - no physical space / time constraints
  - no supervision cost
  - low overhead in finding appropriate volunteers
  - low cost in participant reward

- **Cons**
  - experiments need to be Internet-accessible
  - no or little expertise requirement

Internet users are not all trustworthy ➡ reliability of data inputs may be questionable
Crowdsourcing for Multimedia QoE Assessment

... with an emphasis on ensuring the reliability of data inputs
What is QoE?

Quality of Experience =

Users’ Satisfaction about A Service (Multimedia Content in this context)
Challenges on QoE Assessment

- How to quantify the QoE of multimedia content efficiently and reliably?

- Approaches
  - objective evaluation => inaccurate
  - subjective evaluation => large overhead
Crowdsourcing and Its Applications to Academic Researches / Sheng-Wei Chen
Drawbacks of Subjective Evaluation

- High economic cost
  - Participant reward
  - Transportation cost

- High labor cost
  - Supervision labor

- Physical space/time constraints
  - Laboratory space (cannot do 1000-ppl experiment unless extremely resourceful)
  - Difficult to find participants doing experiments at 3am

→ an appropriate case for crowdsourcing!
Crowdsourcing Challenges

- Not every Internet user is trustworthy
- Experiments without supervision ➔ users may give erroneous feedback perfunctorily, carelessly, or dishonestly
- Increase the variance of the evaluation results and lead to biased conclusions

- Need to have an ONLINE algorithm to detect problematic inputs!
Traditional Approach: Absolute Categorial Rating

- Excellent?
- Good?
- Fair?
- Poor?
- Bad?

Vote

Vote

Crowdsourcing and Its Applications to Academic Researches / Sheng-Wei Chen
Our Proposal: Paired Comparison Test

Which one is better?

Vote

A
Features of Paired Comparison

- Generalizable across a variety of multimedia applications
- **Simple** comparative judgment
- The users’ inputs can be verified
- Interval scale QoE scores can be extracted
Verification of Users’ Inputs

- Transitivity property
  - If \( A > B \) and \( B > C \) \( \Rightarrow \) \( A \) should be \( > C \)

- Transitivity Satisfaction Rate (TSR)
  \[
  \text{TSR} = \frac{\text{# of triples satisfy the transitivity rule}}{\text{# of triples the transitivity rule may apply to}}
  \]

- Detect inconsistent judgments from problematic users
  - \( \text{TSR} = 1 \) \( \Rightarrow \) perfect consistency
  - \( \text{TSR} \geq 0.8 \) \( \Rightarrow \) generally consistent
  - \( \text{TSR} < 0.8 \) \( \Rightarrow \) judgments are inconsistent
Paired Comparison: An Example

- Absolute choice frequencies of 234 subjects choosing between pairs of nine celebrities.
- L. B. Johnson (LBJ), Harold Wilson (HW), and Charles De Gaulle (CDG) are politicians; Johnny Unitas (JU), Carl Yastrzemski (CY), and A. J. Foyt (AJF) are athletes; Brigitte Bardot (BB), Elizabeth Taylor (ET), and Sophia Loren (SL) are female movie stars.

<table>
<thead>
<tr>
<th></th>
<th>LBJ</th>
<th>HW</th>
<th>CDG</th>
<th>JU</th>
<th>CY</th>
<th>AJF</th>
<th>BB</th>
<th>ET</th>
<th>SL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBJ</td>
<td>0</td>
<td>159</td>
<td>163</td>
<td>175</td>
<td>183</td>
<td>179</td>
<td>173</td>
<td>160</td>
<td>142</td>
</tr>
<tr>
<td>HW</td>
<td>75</td>
<td>0</td>
<td>138</td>
<td>164</td>
<td>172</td>
<td>160</td>
<td>156</td>
<td>122</td>
<td>122</td>
</tr>
<tr>
<td>CDG</td>
<td>71</td>
<td>96</td>
<td>0</td>
<td>145</td>
<td>157</td>
<td>138</td>
<td>140</td>
<td>122</td>
<td>120</td>
</tr>
<tr>
<td>JU</td>
<td>59</td>
<td>70</td>
<td>89</td>
<td>0</td>
<td>176</td>
<td>115</td>
<td>124</td>
<td>86</td>
<td>61</td>
</tr>
<tr>
<td>CY</td>
<td>51</td>
<td>62</td>
<td>77</td>
<td>58</td>
<td>0</td>
<td>77</td>
<td>95</td>
<td>72</td>
<td>61</td>
</tr>
<tr>
<td>AJF</td>
<td>55</td>
<td>74</td>
<td>96</td>
<td>119</td>
<td>157</td>
<td>0</td>
<td>134</td>
<td>92</td>
<td>71</td>
</tr>
<tr>
<td>BB</td>
<td>61</td>
<td>78</td>
<td>94</td>
<td>110</td>
<td>139</td>
<td>100</td>
<td>0</td>
<td>67</td>
<td>48</td>
</tr>
<tr>
<td>ET</td>
<td>74</td>
<td>112</td>
<td>112</td>
<td>148</td>
<td>162</td>
<td>142</td>
<td>167</td>
<td>0</td>
<td>87</td>
</tr>
<tr>
<td>SL</td>
<td>92</td>
<td>112</td>
<td>114</td>
<td>173</td>
<td>173</td>
<td>163</td>
<td>186</td>
<td>147</td>
<td>0</td>
</tr>
</tbody>
</table>

![Paired Comparison Chart]
Inference of QoE Scores

- Apply Bradley-Terry-Luce (BTL) model to extract an interval-scale score for each quality level
  - $n$ quality levels: $T_1, \ldots, T_n$
  - $P_{ij}$: the probability of choosing $T_i$ over $T_j$
    $$P_{ij} = \frac{\pi(T_i)}{\pi(T_i) + \pi(T_j)} = \frac{e^{u(T_i) - u(T_j)}}{1 + e^{u(T_i) - u(T_j)}}$$
  - $u(T_i)$ is the estimated QoE score of the quality level $T_i$
Experiment Design

Suppose our task is to evaluate the effect of $n$ audio processing algorithms (e.g., encoding)

1. Select an audio clip (source clip) as the evaluation target
2. Apply the $n$ algorithms to the source clip and generate $n$ different versions of the clip (test clips)
3. Create an Adobe Flash-based system for users to evaluate the $n$ test clips
4. A user need to perform $\binom{n}{2}$ paired comparisons

and not reward a user if his inputs are not self-consistent (i.e., TSR is lower than a certain threshold)
Concept Flow of Acoustic QoE Evaluation

Ready to Make a Decision

Quality with Space Key Released is Better

Vote

Quality with Space Key Pressed is Better
Acoustic QoE Evaluation

- Which one is better?
  - Simple pair comparison

(SPACEx key released)  (SPACEx key pressed)
Optical QoE evaluation

- Which one is better?

(SPACExkey released)  (SPACExkey pressed)
Acoustic QoE Evaluation

- **MP3 compression level**
  - Source clips: one fast-paced and one slow-paced song
  - MP3 CBR format with 6 bit rate levels: 32, 48, 64, 80, 96, and 128 Kbps
  - 127 participants and 3,660 paired comparisons

- **Effect of packet loss rate on VoIP**
  - Two speech codecs: G722.1 and G728
  - Packet loss rate: 0%, 4%, and 8%
  - 62 participants and 1,545 paired comparisons
Evaluation Results

MP3 Compression Level  VoIP Packet Loss Rate

![Graph showing MP3 Compression Level](image)

![Graph showing VoIP Packet Loss Rate](image)
Optical QoE Evaluation

- **Video codec**
  - Source clips: one fast-paced and one slow-paced video clip
  - Three codecs: H.264, WMV3, and XVID
  - Two bit rates: 400 and 800 Kbps
  - 121 participants and 3,345 paired comparisons
Optical QoE Evaluation

- Loss concealment scheme
  - Source clips: one fast-paced and one slow-paced video clip
  - Two concealment schemes
    - Frame copy (FC): conceal errors in a video frame by replacing a corrupted block with the block in the corresponding position in the previous frame
    - Frame copy with frame skip (FCFS): a frame will be dropped if the percentage of corrupted slices in it exceeds 10%; otherwise apply the FC method to conceal the errors
  - Packet loss rate: 1%, 5%, and 8%
  - 91 participants and 2,745 paired comparisons
Evaluation Results

Video Codec

Concealment Scheme

Crowdsourcing and Its Applications to Academic Researches / Sheng-Wei Chen
Participant Source

- Laboratory
  - Recruit part-time workers at an hourly rate of 8 USD

- MTurk
  - Post experiments on the Mechanical Turk web site
  - Pay the participant 0.15 USD for each qualified experiment

- Community
  - Seek participants on the website of an Internet community with 1.5 million members
  - Pay the participant an amount of virtual currency that was equivalent to one US cent for each qualified experiment
Participant Source Evaluation

- With crowdsourcing...
  - lower monetary cost
  - wider participant diversity
  - maintaining the evaluation results’ quality

<table>
<thead>
<tr>
<th>Case Study</th>
<th>Experimenter Source</th>
<th>Total Cost (dollar)</th>
<th># Rounds</th>
<th># Person</th>
<th>Qualified Rate</th>
<th>Cost / Round (cent)</th>
<th>Time / Round (sec)</th>
<th>Avg. TSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP3 Bit Rate</td>
<td>Laboratory</td>
<td>50.97</td>
<td>1440</td>
<td>10</td>
<td>67%</td>
<td>3.54</td>
<td>16</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>MTurk</td>
<td>7.50</td>
<td>750</td>
<td>24</td>
<td>47%</td>
<td>1.00</td>
<td>9</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Community</td>
<td>1.03</td>
<td>1,470</td>
<td>93</td>
<td>54%</td>
<td>0.07</td>
<td>25</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Crowdsourcing is an ideal strategy for multimedia QoE assessment!

Quadrant of Euphoria

http://mmnet.iis.sinica.edu.tw/link/qoe
Conclusion

- Crowdsourcing provides a new paradigm and a new platform for academic researches

- New applications, new methodologies, and new businesses are emerging with the advent of crowdsourcing technology
謝謝各位！

陳昇瑋
中央研究院 資訊科學研究所

http://www.iis.sinica.edu.tw/~swc