Efficient Crowdsourcing for Metadata Generation

Wolf-Tilo Balke

Institute for Information Systems
Technische Universität Braunschweig
http://www.ifis.cs.tu-bs.de
Most semantic retrieval tasks still rely on metadata

- But how to know what metadata will be needed at content creation time?
- Can missing metadata be efficiently created just in time also for unexpected queries?

Joint work with Joachim Selke & Christoph Lofi
Crowdsourcing

• Hot and emerging paradigm
  – Vaguely defined concept: “Concepts for fostering human collaboration to solve complex problems.”
  – Aims at tapping the “the Wisdom of the Crowd”
    • “Under certain conditions large crowds of people are able to perform highly effective decisions”
Generic Crowd-sourcing

- **Generic Task-Based Crowd-sourcing**
  - General purpose platforms can facilitate virtually any task for anybody
  - Workers are attracted and retained by **paying money**
Crowd-Enabled Databases

- Core idea: Build a **database engine** which can dynamically crowdsource certain operations
  - **Complete missing** data during query time
    - Incomplete tuples (CNULL values)
    - Elicit completely new tuples
  - **Use human intelligence operators**
    - Entity resolution
    - Similarity rankings
    - etc.

```sql
CREATE TABLE Department ( 
  university STRING, 
  name STRING, 
  url CROWD STRING, 
) 

SELECT market_capitalization FROM company WHERE name = "I.B.M.";
```
So, how about Metadata Generation?

• The ease-of-use and reliability of crowdsourcing tasks *varies* with the respective use case
• In general, three variables have to be controlled
  – **Answer/Solution Quality**, impacted by…
    • Worker diligence
    • Worker maliciousness
    • Worker quality and skills
  – **Execution Time**
    • Job attractiveness (payment vs. time)
    • Worker pool size
  – **Costs**
    • Number of HITs
    • costs per HIT (affected by time and skill needed)
    • Quality control overhead
Crowdsourcing in Action

• Popular example from art: Aaron Koblin
  – Laboral Centro de Arte, Gijon, Spain
  – Japan Media Arts Festival, Tokyo, Japan
  – Apex Gallery, New York, USA
  – ElectroFringe, New Castle, Australia
  – Media Art Friesland, The Netherlands
Crowdsourcing in Action

• You get what you pay for…
  – 10 000 sheep = 200 USD
Crowdsourcing in Action

Draw a sheep facing to the left

$0.02

BAAA!
• **How to perform better?**
  – Employ hybrid techniques combining crowdsourcing, information extraction, machine learning and the Social Web!

• **Tackle the following challenges**
  – **Performance**
    • Drastically speed up crowdsourcing times
  – **Costs**
    • Require just few crowdsourcing HITs for obtaining a large number of judgements
  – **Data Quality**
    • Circumvent the impact of malicious workers
    • Reliably obtain judgements for even obscure and rare items
• **Reconsider crowd-enabled databases**
  
  – Large table with movies
    * e.g. like IMDb, ~2 Million movies

  – **Task**
    * Introduce new column with a rating for humour (0-10)

  – **Traditional approach**
    * Create crowd-sourcing task asking users for judgement
    * Consensual result requiring background knowledge

  – **Extremely challenging (and expensive) task!**

---

*Pushing the Boundaries of Crowd-enabled Databases*
• The **Social Web** as a Data Source has become common-place
  – Collect information before buying products (reviews)
  – Recommend news articles, movies, books,…

• Mostly all this data is aggregated into a **rating**
  – Easy to do, rich in information, and rather ubiquitous
  – Valuable to extract: collaborative filtering, etc.
• **Idea:**
  – Each user has personal likes/dislikes, preferences, etc. that explain the respective rating behaviour
  – Ratings of each individual will be rather consistent regarding likes/dislikes... a **systematic bias**
  – How to **dissemble** ratings into the individual biases?

• **Let users and items be d-dimensional points**
  – Coordinates of a user represent his/her personality (bias)
  – Coordinates of an item represent its profile regarding personality traits
• Building the perceptual space
  – Possible from ratings, review texts, tags,…

• Factor Models
  – Developed to estimate the value of non-observed ratings for the purpose of recommending new unrated items
  – Ratings are seen as a function of user vectors and item vectors
  – Prominent factor models: SVD, Euclidian embedding,…
Perfect Use case

- The Godfather
- Rambo
- Shooter
- Behind Enemy Lines
- Finding Nemo
- Toy Story
- Kung Fu Panda
- Shrek
- Chronicles of Narnia
- Star Wars: Episode VI
- Star Wars: Episode I
- Star Trek

- very humorous (10-8)
- humorous (7-5)
- some humor (4-3)
- grave (2-1)
How to use a Perceptual Space?

• Extract the **correct distances** regarding the topic of interest from the perceptual space…
  – However, the data is hidden in the space!
  – What dimensions should contribute to the distances?

• **Main idea:** train a classifier via crowdsourcing
  – Provide training set via the crowd: positive and negative examples for humorous movies, good books,…
  – Non-linear SVM for classification
  – Non-linear regression for values
How to use a Perceptual Space?

The Social Web
- Tags
- Reviews
- Ratings
- Links

Extract

Perceptual space

Attribute values

Responses

Crowd-enabled DB

Query

Result

HITs

Crowdsourcing service
• Perceptual spaces compared to Metadata
  – LSI over all available movie metadata
  – g-means for $n$ positive and $n$ negative examples

<table>
<thead>
<tr>
<th>Genre</th>
<th>Random</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 40$</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 40$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Netflix</td>
</tr>
<tr>
<td>Comedy</td>
<td>0.50</td>
<td>0.58</td>
<td>0.70</td>
<td>0.76</td>
<td>0.46</td>
<td>0.30</td>
<td>0.28</td>
<td>0.85</td>
</tr>
<tr>
<td>Documentary</td>
<td>0.50</td>
<td>0.73</td>
<td>0.81</td>
<td>0.84</td>
<td>0.64</td>
<td>0.63</td>
<td>0.62</td>
<td>0.96</td>
</tr>
<tr>
<td>Drama</td>
<td>0.50</td>
<td>0.60</td>
<td>0.66</td>
<td>0.73</td>
<td>0.51</td>
<td>0.45</td>
<td>0.49</td>
<td>0.86</td>
</tr>
<tr>
<td>Family</td>
<td>0.50</td>
<td>0.82</td>
<td>0.86</td>
<td>0.88</td>
<td>0.44</td>
<td>0.43</td>
<td>0.43</td>
<td>0.95</td>
</tr>
<tr>
<td>Horror</td>
<td>0.50</td>
<td>0.83</td>
<td>0.86</td>
<td>0.87</td>
<td>0.53</td>
<td>0.32</td>
<td>0.43</td>
<td>0.92</td>
</tr>
<tr>
<td>Romance</td>
<td>0.50</td>
<td>0.56</td>
<td>0.68</td>
<td>0.73</td>
<td>0.45</td>
<td>0.35</td>
<td>0.38</td>
<td>0.91</td>
</tr>
<tr>
<td>Mean</td>
<td>0.50</td>
<td>0.69</td>
<td>0.76</td>
<td>0.80</td>
<td>0.50</td>
<td>0.41</td>
<td>0.44</td>
<td>0.91</td>
</tr>
</tbody>
</table>
• Discussion of **crowdsourcing** for just in time metadata generation
  – **Quality** of crowdsourcing tasks needs to be addressed
    • Correctness, time, and costs
    • What type of task, possible quality assurance,…
  – Training classifiers over **perceptual spaces** can solve the problem to some degree
References


• Franklin, M., Kossmann, D., Kraska, T., Ramesh, S., Xin, R.: *CrowdDB: Answering Queries with Crowdsourcing*. ACM SIGMOD Int. Conf. on Management of Data, Athens, Greece, 2011.

• Selke, J., Lofi, C., Balke, W.-T.: *Pushing the Boundaries of Crowd-Enabled Databases with Query-Driven Schema Expansion*. 38th Int. Conf. on Very Large Data Bases (VLDB), in PVLDB 5(2), Istanbul, Turkey, 2012.

Thanks for Your Attention

Questions?

balke@ifis.cs.tu-bs.de