Crowdsourcing Feedback for Pay-As-You-Go Data Integration

Fernando Osorno-Gutierrez, Norman W. Paton, Alvaro A. A. Fernandes
School of Computer Science, University of Manchester
CONTEXT AND MOTIVATION
Context

● Classical data integration
  – Heavy upfront investment.
  – Costly to rely on experts, costly to wait months till deployment.
  – High initial result quality.
  – Degradation leading to costly maintenance when sources change fast.

● Pay-as-you-go data integration
  – Very low upfront investment.
  – Automate the bootstrapping (i.e., using model management techniques).
  – Quick deployment, but low initial result quality.
  – Improvement through feedback, maintenance is integral.
  – A platform for this kind of integration is referred to as a Dataspace.
Goals > Questions

- Best-effort initial results from bootstrapping motivate users to "pay" in the form of feedback.
- Make the most of the feedback obtained so as to keep "return on investment" high for users.

- Can we make good use of feedback from non-experts?
- How much feedback do we need?
- How reliable is that feedback?
Motivation

- [Belhajjame et al.]
  - Feedback on query results
  - End (i.e., non-expert) user states whether a result tuple is a true or a false positive.
  - Main contributions were:
    - Using this feedback to select and refine algorithmically-generated mappings
    - Showing that around 150 feedback instances are enough to bring the F-measure-combined precision and recall close to 100%
  - But... uses synthetic feedback.

- What would these results be like if we used feedback from humans through crowdsourcing?

Motivation

- **[Belhajjame et al.]**
  - Feedback on query results
  - End (i.e., non-expert) user states whether a result tuple is a true or a false positive.
  - Main contributions were:
    - Using this feedback to select and refine algorithmically-generated mappings
    - Showing that around 150 feedback instances are enough to bring the F-measure-combined precision and recall close to 100%
  - But... uses synthetic feedback.

- **What would these results be like if we used feedback from humans through crowdsourcing?**

RESEARCH ISSUES AND APPROACH
The Data Integration Problem

Find information about “The Beatles”

Global Schema

Source Schema 1
Source 1

Source Schema 2
Source 2

Source Schema 3
Source 3

What are the correct results?
Dataspace Lifecycle/Initialisation

We'll focus here, in this talk.

We'll focus here, in this talk.

We'll first show how we created a case study here.

CASE STUDY
A Case Study in Bootstrapping

**Source Schema**
- `musicbrainz_artist`
  - name
  - gender
  - country
  - type
  - begin_date_year

- `discogs_artist`
  - name
  - realname

**Global Schema**
- `artist`
  - name
  - realname
  - gender
  - country
  - type
  - begin_date_year

**Musicbrainz**
- `artist`

**Discogs**
- `artist`
Generation of Schema Mappings with Spicy

- Spicy* generated candidate schema mappings.

- We selected 3 schema mappings (M1, M2 and M3) with the attributes name, country and type.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Source schema attribute</th>
<th>Global schema attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>discogs_artist.name</td>
<td>name</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.country</td>
<td>country</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.type</td>
<td>type</td>
</tr>
</tbody>
</table>

Generation of Schema Mappings with Spicy

- Spicy* generated candidate schema mappings.

- We selected 3 schema mappings (M1, M2 and M3) with the attributes name, country and type.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Source schema attribute</th>
<th>Global schema attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>discogs_artist.name</td>
<td>name</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.country</td>
<td>country</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.type</td>
<td>type</td>
</tr>
</tbody>
</table>

Generation of Schema Mappings with Spicy

- Spicy* generated candidate schema mappings.

- We selected 3 schema mappings (M1, M2 and M3) with the attributes name, country and type.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Source schema attribute</th>
<th>Global schema attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>discogs_artist.name</td>
<td>name</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.country</td>
<td>country</td>
</tr>
<tr>
<td></td>
<td>musicbrainz_artist.type</td>
<td>type</td>
</tr>
</tbody>
</table>

Obtaining Mappings with Varying Levels of Correctness

- We know that, in practice, different sources have instance data with different quality regarding correctness.
- We therefore copied the results from M1.
- We changed a percentage of the country column to randomly selected incorrect values.
- This gave rise to sets of results corresponding to the hypothetical mappings M4-M7 below.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Percentage of tuples modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>M4</td>
<td>20%</td>
</tr>
<tr>
<td>M5</td>
<td>40%</td>
</tr>
<tr>
<td>M6</td>
<td>60%</td>
</tr>
<tr>
<td>M7</td>
<td>80%</td>
</tr>
</tbody>
</table>
Recall that in [Belhajjame et al.], feedback is syntactic.

No issues of inconsistency response arise: the algorithms used are deterministic.

With humans, we need to consider that:

- The same user can give two different answers to the same question if asked twice.
- Are the answers given by user U reliable?
- Two different users can give two different answers to the same question.
- Are the answers given by users U reliable when compared with the answers given by user U'?
User Reliability

- We use controlled redundancy to measure reliability.
- We then measure the effect of filtering out feedback from unreliable users.
- Intra-observer reliability (IaOR) is the level of consistency of a given respondent.
  - Using replication, for some of the questions, we ask the same respondent the same question at two different times.
  - We deem respondents are reliable if they provide the same answer to a replicated question both times.
- Inter-observer reliability (IrOR) is the level of consistency of a given respondent relative to other respondents.
  - Using replication, for some of the questions, we ask two respondents the same question.
  - We deem respondents reliable if they give the same answers as other users.
IMPLEMENTATION
Sampling Mapping Results

- Each mapping produced results.

Results M1
(~4K tuples)

Population M1
(~4K tuples)

Sample M1
(~10%)  
(times 7 mappings)

Distribution of tuples

Questionnaires
Help us to evaluate information about music artists

Show instructions

Question 1. Is the following information about a music artist correct?

Artist name: Green Day
Country: United States
Type of artist: Group

- Correct
- Incorrect

Next

Submit
Crowdsourcing application

Characteristics:
- Type of HIT: External Questions.

Capabilities:
- Managed the HITs and the estimation of the reliability.
- Hosted HITs.
Distribution of Tuples

Remember that the tuples from the samples of tuples had to be distributed into questionnaires.

Special requirements of one questionnaire:

- No tuples about the same artist.
- Minimize the number of tuples from the same mapping.
- Questionnaires generated after distribution: ~100
- A questionnaire contained 25 unique tuples.
- Each questionnaire featured in one HIT.
Intra Observer Reliability (IaOR)

- Groups of 2 HITs.
- We copy 3 questions.
- HITs are answered by the same respondent 2 hours apart.
Inter Observer Reliability (IrOR)

- Groups of 3 HITs.
- We copy 2 questions.
- HITs are answered by different respondents.
EVALUATION
Error in Precision per Mapping
Error in Precision per Mapping

Error down to 0.1 after ~50
Average Error in Precision (synthetic feedback, Belhajjame et al.)

Fig. 2. Average error in precision.

Fig. 3. Average error in recall.

We chose to run the experiment for 25 times because it has been shown statistically to yield good estimations [54].
consider tuple-based feedback. Specifically, we randomly generated a stratified sample of 10 annotated tuples out of the set of tuples returned by the candidate mappings. Stratified sampling ensures that the members of a population are first grouped into relatively homogeneous sub-populations before sampling. Stratified sampling is performed as follows. Given the ground truth which partitions the tuples retrieved by the candidate mappings into expected and unexpected, a sample of 10 tuples is randomly selected from that set of tuples retrieved by the candidate mappings, and for which feedback has not been provided yet, such that the ratio of expected (resp. unexpected tuples) in the sample is the same as the ratio of expected (resp. unexpected) tuples within the ground truth. This sampling method improves the representativeness of the feedback instances generated. It ensures that the number of expected and unexpected tuples in the sample is proportional to the number of expected and unexpected tuples in the result set obtained using all candidate mappings.

To measure the quality of the relative precision and recall, we repeated the annotation experiment described above multiple times, specifically 25 times, and computed the average error in precision and recall at every feedback iteration. The error in precision (resp. recall) is the difference between the relative precision (resp. recall) of a candidate mapping computed using supplied feedback and the "ground truth" precision (resp. recall). Fig. 2 illustrates the average error in precision, and Fig. 3 illustrates the average error in recall. Note that the scale of the vertical axis is different in the two figures. The two figures show that when the user provided 10 feedback instances the error in both precision and recall are relatively high. For example, the error in precision is 0.23 and the error in recall is 0.22 for the mappings that

9 We chose to run the experiment for 25 times because it has been shown statistically to yield good estimations [54].

Error down to 0.1 after ~70
Consider tuple-based feedback. Specifically, we randomly generated a stratified sample of 10 annotated tuples out of the set of tuples returned by the candidate mappings. Stratified sampling ensures that the members of a population are first grouped into relatively homogeneous sub-populations before sampling. Stratified sampling is performed as follows. Given the ground truth which partitions the tuples retrieved by the candidate mappings into expected and unexpected, a sample of 10 tuples is randomly selected from that set of tuples retrieved by the candidate mappings, and for which feedback has not been provided yet, such that the ratio of expected (resp. unexpected tuples) in the sample is the same as the ratio of expected (resp. unexpected) tuples within the ground truth. This sampling method improves the representativeness of the feedback instances generated. It ensures that the number of expected and unexpected tuples in the sample is proportional to the number of expected and unexpected tuples in the result set obtained using all candidate mappings.

To measure the quality of the relative precision and recall, we repeated the annotation experiment described above multiple times, specifically 25 times, and computed the average error in precision and recall at every feedback iteration. The error in precision (resp. recall) is the difference between the relative precision (resp. recall) of a candidate mapping computed using supplied feedback and the "ground truth" precision (resp. recall). Fig. 2 illustrates the average error in precision, and Fig. 3 illustrates the average error in recall. Note that the scale of the vertical axis is different in the two figures. The two figures show that when the user provided 10 feedback instances the error in both precision and recall are relatively high. For example, the error in precision is 0.23 and the error in recall is 0.22 for the mappings that received 10 feedback instances.

We chose to run the experiment for 25 times because it has been shown statistically to yield good estimations [54].
Average Error in Precision

![Graph showing the average error in precision with different feedback amounts.](image)

- Blue line: No filter
- Red dotted line: Intra observer reliability
- Green line: Inter observer reliability
- Purple dotted line: Inter and intra observer reliability

Feedback Amount

DBCrowd 2013
Average Error in Precision

Error down to 0.1 after ~20
Average Error in Precision

Error down to 0.1 after ~20

Stable at 0.04 or less after ~250
Average Error in Precision

Error down to 0.1 after ~20

Stable at 0.04 or less after ~250

Wait for an explanation of this inflection!
Average Error in Precision randomised

- No filter
- Intra observer reliability
- Inter observer reliability
- Inter and Intra observer reliability

Feedback Amount

Error in Precision
Average Error in Precision randomised

Previous plot reflected sequence of feedback arrival!
Average Error in Precision randomised

To generalize, we shuffle the sequence to obtain thousands of randomizations.

Previous plot reflected sequence of feedback arrival!
To generalize, we shuffle the sequence to obtain thousands of randomizations. This gives smoother curves with no inflection. Previous plot reflected sequence of feedback arrival!
To generalize, we shuffle the sequence to obtain thousands of randomizations. This gives smoother curves with no inflection.

Error down to 0.1 after ~15
To generalize, we shuffle the sequence to obtain thousands of randomizations. This gives smoother curves with no inflection.

Previous plot reflected sequence of feedback arrival.

Error down to 0.1 after ~15
Stable at less than 0.04 after ~ 200
Average Error in Precision randomised

Previous plot reflected sequence of feedback arrival!

To generalize, we shuffle the sequence to obtain thousands of randomizations.

This gives smoother curves with no inflection.

Error down to 0.1 after ~15
Stable at less than 0.04 after ~200

There is benefit in filtering for IaOR and IrOR.
Conclusions

- We have confirmed the conclusions in Belhajjame et al. that modest amounts of feedback on query results by non-experts can lead to significant improvement in precision.

- We have found that crowdsourcing feedback leads to better results than hypothesized by Belhajjame et al. on the basis of synthetic feedback.

- We have described and implemented techniques for filtering feedback on the basis of intra- and inter-observer reliability.

- We have shown this filtering to improve quality.
Questions
References


Appendix A: Annotating Mappings

- Measures used in Belhajjame et al. [1].
- Precision

\[
\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
\]

- Error in Precision

\[
\text{Error in Precision} = |\text{GTP} - \text{Precision}|
\]

Where:
- \textit{GTP}: Ground Truth Precision
- \textit{Precision}: Precision estimated using user feedback.
Appendix B: Average Error in Precision

- Average Error in Precision (AEP)

\[
\text{Average Error in Precision}_{i} = \frac{\sum_{j=1}^{K} |GTP_j - \text{Precision}_{ij}|}{K}
\]

Where:
- \(K\): Number of mappings
- \(GTP_j\): Ground Truth of precision of the mapping \(j\).
- \(\text{Precision}_{ij}\): Value of precision after \(i\) feedback instances
- \(i\): Current number of evaluation after \(i\) feedback instances