Automated Selection of Multiple Datasets for Extension by Integration

Yael Amsterdamer  Moran Ben-Yehuda
Extension by Integration

- reviews
- vendors
- manufacturers
- parts
Scenario

• We have an initial data table (e.g., csv)
• We want to extend this table by integration with other sources
• Which ones to choose?
  • Amount of added data
  • Introduced errors
  • Completeness
  • Quality of matching to the initial table
• A greater challenge: integrating multiple tables
## Integration Result: Products & CA

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>NULL</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>NULL</td>
</tr>
</tbody>
</table>

## CA (Companies in Africa)

<table>
<thead>
<tr>
<th>Company</th>
<th>Located</th>
<th>Category</th>
<th>Rev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCnD</td>
<td>NULL</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Macron</td>
<td>NULL</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Transact</td>
<td>Senegal</td>
<td>Finance</td>
<td>87.6</td>
</tr>
<tr>
<td>XYnZ</td>
<td>Tunisia</td>
<td>NULL</td>
<td>252.2</td>
</tr>
</tbody>
</table>

## MEIT (Middle Eastern IT)

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macron</td>
<td>Egypt</td>
<td>155</td>
</tr>
<tr>
<td>Netter</td>
<td>UAE</td>
<td>32</td>
</tr>
<tr>
<td>Opportune</td>
<td>Qatar</td>
<td>79</td>
</tr>
<tr>
<td>Promot</td>
<td>Israel</td>
<td>35</td>
</tr>
<tr>
<td>QueenTech</td>
<td>Jordan</td>
<td>28</td>
</tr>
</tbody>
</table>

### Automated Selection of Multiple Datasets for Extension by Integration / Y. Amsterdamer, M. Ben-Yehuda
### Integration Result: Products & CA & MEIT

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>NULL</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
</tbody>
</table>
Our two main problems [CIKM 2021]:

1. Define a **metric** for the “value” of integration

2. **Efficiently find** the subset of relations that maximizes it
Previous Work: Source Selection

Finding links between relations
Domain Search
Finding joinable/unionable relations

Source Selection for Data Fusion

Metrics for table join in interactive data science (Zhang et al. 2020)

Data augmentation for machine learning (Chepurko et al. 2020)
Outline

• Problem definition
  ➢ based on integration gain and cost
• Algorithms
• Experimental study
Metrics for Valuable Integration

• By properties of the integration result
  
  Some derivable from
  → Properties of integrated relations
  → Quality of integration

• Cost and gain of the integration
Let us start from the end:

• Assume a black-box for multi-relation integration

  • **Integration gain** number of correct values in the integration result
    • Expected
    • How do we compute correctness likelihood?

• Integration cost:
  • **Incompleteness cost** number of NULLs
  • **Error cost** expected number of erroneous values
  • **Fixed cost per integration**

If the black box provides cell correctness probability estimation – we are done.
Properties of Integrated Relations

- Initial relation $R_0$, set of candidate relations $\mathcal{R} = \{R_1, R_2, \ldots\}$
- Each $R_i$ has
  - $U^i = \{U^i_1, U^i_2, \ldots\}$ - attributes
  - $\text{key}(R_i)$
  - Tuples with values (possibly NULL)
  - $p_{\text{correct}}(R_i)$ - probability of error in each value

Why compute this for the input and not the output?
Properties of Integration

• Many existing tools for data integration
  • Matching attributes
  • linking records
  • Mostly for relation pairs

• The integration result is

\[ R = \mathcal{I}_{\text{res}}(... \mathcal{I}_{\text{res}}(\mathcal{I}_{\text{res}}(R_0, R_{i_1}), R_{i_2})..., R_{i_m}) \]
Correctness Derivation: Linking tuples

• Let $t$ be a tuple matched to tuple $t'$ based on their values in $U, U'$
  • E.g., GreatPad X4000 matched to BCnD based on their values on attributes Manu. and Company

• Link weight for tuple $t$:

$$W_{\text{link}}(t) := \text{pattMatch}(U) \cdot W_{\text{correct}}(U, t) \cdot W_{\text{correct}}(U', t') \cdot \text{pvalMatch}(t(U), t'(U'))$$

- Attributes are matched correctly
- ...and value in left table is correct
- ...and value in right table is correct
- ...and values indeed match
## Products

<table>
<thead>
<tr>
<th>Prod</th>
<th>Manu.</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>NULL</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
</tr>
</tbody>
</table>

## CA (Companies in Africa)

<table>
<thead>
<tr>
<th>Company</th>
<th>Located</th>
<th>Category</th>
<th>Rev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCnD</td>
<td>NULL</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Macron</td>
<td>NULL</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Transact</td>
<td>Senegal</td>
<td>Finance</td>
<td>87.6</td>
</tr>
<tr>
<td>XYnZ</td>
<td>Tunisia</td>
<td>NULL</td>
<td>252.2</td>
</tr>
</tbody>
</table>

## MEIT (Middle Eastern IT)

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macron</td>
<td>Egypt</td>
<td>155</td>
</tr>
<tr>
<td>Netter</td>
<td>UAE</td>
<td>32</td>
</tr>
<tr>
<td>Opportune</td>
<td>Qatar</td>
<td>79</td>
</tr>
<tr>
<td>Promot</td>
<td>Israel</td>
<td>35</td>
</tr>
<tr>
<td>QueenTech</td>
<td>Jordan</td>
<td>28</td>
</tr>
</tbody>
</table>

## Integration Result: Products & CA & MEIT

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>Egypt</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
</tbody>
</table>
Correctness Derivation: Values

• Correctness of $t(U)$ in $R$
  • If $R$ is an input relation, $W^{\text{correct}}(U, t) := P^{\text{correct}}(R_i)$
  • If $t(U)$ was added in a new column as a result of integrating $R'$,
    $W^{\text{correct}}(U, t) := W^{\text{link}}(t) \cdot W^{\text{correct}}(U, t')$
  • If $t(U)$ was a NULL resolved by integrating $R'$,
    $W^{\text{correct}}(U, t) := W^{\text{link}}(t) \cdot P^{\text{attMatch}}(U, U') \cdot W^{\text{correct}}(U', t')$
### Products

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>Egypt</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td>NULL</td>
<td>32</td>
</tr>
</tbody>
</table>

### CA (Companies in Africa)

<table>
<thead>
<tr>
<th>Company</th>
<th>Located</th>
<th>Category</th>
<th>Rev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCnD</td>
<td>NULL</td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Macron</td>
<td>NULL</td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Insact</td>
<td>Senegal</td>
<td>Finance</td>
<td>87</td>
</tr>
<tr>
<td>Netter</td>
<td>UAE</td>
<td>NULL</td>
<td>32</td>
</tr>
</tbody>
</table>

### MEIT (Middle Eastern IT)

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macron</td>
<td>Egypt</td>
<td>155</td>
</tr>
<tr>
<td>Netter</td>
<td>UAE</td>
<td>32</td>
</tr>
<tr>
<td>XYnZ</td>
<td>Tunisia</td>
<td>252.2</td>
</tr>
</tbody>
</table>

### Integration Result: Products & CA

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GreatPad X4000</td>
<td>BCnD</td>
<td>South Africa</td>
<td></td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>GreatPad Y6000</td>
<td>BCnD</td>
<td>South Africa</td>
<td></td>
<td>Technology</td>
<td>115.8</td>
</tr>
<tr>
<td>Superb Vital</td>
<td>Macron</td>
<td>Egypt</td>
<td></td>
<td>IT</td>
<td>155.3</td>
</tr>
<tr>
<td>Smarterbook Elite</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td></td>
<td>NULL</td>
<td>32</td>
</tr>
<tr>
<td>Smarterbook Emerge</td>
<td>Netter</td>
<td>Saudi Arabia</td>
<td></td>
<td>NULL</td>
<td>32</td>
</tr>
</tbody>
</table>

*Automated Selection of Multiple Datasets for Extension by Integration / Y. Amsterdamer, M. Ben-Yehuda*
Formal Pairwise Definitions

**Gain** expected number of correct values.

\[
gain(R_i, R_j) := \sum_{t \in \mathcal{I}_{res}(R_i, R_j)} \sum_{U \in U_{i,j}, t(U) \neq NULL} W^{correct}(U, t)
\]

**Integration cost:**

- **Incompleteness cost** number of NULL in the integration result

\[
\text{Cost}_{\text{NULL}}(\mathcal{I}_{res}(R_i, R_j))
\]

- **Error cost** expected number of erroneous values.

\[
\text{Cost}_{\text{err}}(\mathcal{I}_{res}(R_i, R_j)) := \sum_{t \in \mathcal{I}_{res}(R_i, R_j)} \sum_{U \in U_{i,j}, t(U) \neq NULL} (1 - W^{correct}(U, t))
\]

- **Fixed cost per integration** e.g., monetary cost

\[
\text{Cost}_{\text{fixed}}(\mathcal{I}_{res}(R_i, R_j)) := \text{Cost}_{\text{fixed}}(R_i) + \text{Cost}_{\text{fixed}}(R_j)
\]
OPT-EXTENSION

Find Sub-sequence $R_{i_1}, R_{i_2}, ..., R_{i_m}$

Integration Result

$$R = J_{\text{res}}(...J_{\text{res}}(J_{\text{res}}(R_0, R_{i_1}), R_{i_2})..., R_{i_m})$$

Maximize metric

$$\text{score}(R, \alpha, \beta, \gamma) := \text{gain}(R) - (\alpha \text{Cost}_{\text{NULL}}(R) + \beta \text{Cost}_{\text{err}}(R) + \gamma \text{Cost}_{\text{fixed}}(R))$$
Hardness of OPT-EXTENSION

• OPT-EXTENSION is $\text{FP}^{\text{NP}}$-hard
  • By a reduction from SET COVER
  • Membership result

• Score function is not monotone / convex
Our Solution Scheme

- Iteratively select the next relation to integrate
  - using function $f$
- Exhaustively integrate
- Select intermediate best result

Automated Selection of Multiple Datasets for Extension by Integration / Y. Amsterdamer, M. Ben-Yehuda
Our Solution Scheme

- Iteratively select the next relation to integrate
  - using function $f$
- Exhaustively integrate
- Select intermediate best result
Our Solution Scheme

• Iteratively select the next relation to integrate
  • using function $f$
• Exhaustively integrate
• Select intermediate best result
Selection Criteria

EDMINT-Greedy:
• Greedily maximize the score at each iteration
• Empirically achieves near-optimal scores
• But: performs many integrations
Selection Criteria

EDMINT-Opt:

- **Reduce number of integrations by**
  - Identifying relations that cannot increase the score
  - Identifying relations whose marginal contribution is fixed

```
Algorithm 2: Edmint-Opt implementation of f (selection of the next relation to integrate in Algorithm 1)

Input: R: initial relation; \( \mathcal{R} = \{R_1, \ldots, R_n\} \): set of candidate relations; \( \mathcal{I} \): integration black-box; \( \alpha, \beta, \gamma \): weights for score function

Output: \( R_{\text{max}} \in \mathcal{R} \): single relation to integrate; \( \mathcal{R} \) updated set of candidates

1. for \( R_i \in \mathcal{R} \) do
   2. maxDelta \(<- \text{gain}(R_i) - \beta \text{Cost}_{\text{err}}(R_i) - \gamma \text{Cost}_{\text{fixed}}(R_i)\):
      3. if maxDelta < 0 then \( \mathcal{R} \leftarrow \mathcal{R} - \{R_i\} \):
      4. \( R_{\text{max}} \leftarrow \emptyset \):
      5. maxScore \(<- -\infty \):
      6. for \( R_i \in \mathcal{R} \) do
         7. if \( \forall U_k \in U^i \), atMatch\(^+\)\(_{R_i, \mathcal{R}}(U_k) = \emptyset \) then
            8. delta \(<- \text{score}(\mathcal{I}_{\text{res}}(R, R_i)) - \text{score}(R)\):
               9. if atMatch\(^-\)\(_{R,R_i}(\text{key}(R_i)) \neq \perp \) and delta > 0 then
                  Return \( R_i, \mathcal{R} \):
               10. else \( \mathcal{R} \leftarrow \mathcal{R} - \{R_i\} \):
            else if atMatch\(^-\)\(_{R,R_i}(\text{key}(R_i)) \neq \perp \) and \( R' \leftarrow \mathcal{I}_{\text{res}}(R, R_i) \):
                Update \( R_{\text{max}}, \text{maxScore} \) if the score of \( R' \) is greater;
      11. return \( R_{\text{max}}, \text{maxScore} \);
```
Selection Criteria

EDMINT-Opt:

- Reduce number of integrations by
  - Identifying relations that cannot increase the score
  - Identifying relations whose marginal contribution is fixed

The maximal marginal contribution of a relation (if all non-NUL cells are added, and there are no added NULLs)

There are no additional matches of this candidate relation to other candidates
Integrations are still a bottleneck

• We use an implementation based on locality sensitive hashing (LSH):
  • Attribute sketches used to estimate matching probability
  • An index used to find matches in constant time
  • Depends on the attribute matching method
Experimental Study
Compared algorithms

- **AccDesc** - $f$ greedily selects the most accurate relation that can be integrated
- **Random** - $f$ selects a random relation that can be integrated
- **Brute-Force**
- **EDMINT-Greedy**
- **EDMINT-Opt**
Metrics

We consider three general types of metrics for the integration result:

• Score of the integration
• Number of rounds
• Number of integrations
Datasets

• **Kaggle Collection**: 40 relations related to movies and books → scenario: user already collected relevant datasets

• **Medley Collection**: 100 relations on various topics → scenario: using a data lake

• Each relation consists of 120-1M tuples and 2-67 attributes.
Experimental Results

Varying candidate collection sizes:
Experimental Results

Comparison to the optimal algorithm:

![Comparison Graphs]

Automated Selection of Multiple Datasets for Extension by Integration / Y. Amsterdamer, M. Ben-Yehuda
Experimental Results

Varying the parameters:
### Execution times

| $|R|$ | $\sum |U^i|$ | % match | Index time | Avg. Integration time |
|----|---------|---------|-----------|-----------------------|
|    |         |         |           | Random                | AccDesc | Edmint-Greedy | Edmint-Opt |
| 90 | 3628    | 0.11    | 07:32     | <00:01               | <00:01  | <00:01       | <00:01     |
| 70 | 2221    | 0.15    | 05:09     | <00:01               | <00:01  | <00:01       | <00:01     |
| 50 | 2358    | 0.13    | 03:33     | <00:01               | <00:01  | <00:01       | <00:01     |
| 19 | 300     | 0.63    | 03:07     | 00:02                | 00:01   | 00:02        | 00:02      |
| 12 | 183     | 0.85    | 00:35     | 00:01                | 00:01   | <00:01       | <00:01     |
| 5  | 76      | 0.4     | 00:22     | <00:01               | <00:01  | <00:01       | 00:01      |
Summary

• We defined the problem of extension by integration
  • Cost and gain of integration
  • Using pair-wise black-boxes for attribute matching and tuple linking
  • Direct optimization is hard

• We proposed a scheme and algorithms for the solution

• Experiments on real data
  • Near-optimal score
  • Efficiency by reducing the number of integrations

• Our solution can be combined with various integration methods
Future work

• Extending non-relational data

• Accounting explicitly for other aspects of integration
  • Relevance
  • Data cleaning
  • Data fusion

• Perform automated transformations (group by, filter, pivot) on relations to improve integration quality
Thank You!