Data cleansing

Juhani Ojares
3.4.2017
Contents

- Introduction
- Traditional methods
- Big data method: KATARA
- Summary
Introduction

• Why data needs cleaning?
  – 25% critical corporate data is dirty
  – Compromises queries and analysis tasks
    • → wrong price data in retail databases alone costs US consumers $2.5 billion annually
Introduction

● What is data cleaning?
  – making the data *consistent*
  – process of *detecting* and *correcting* errors in data

● What is dirty data?
  – Duplicates
  – Miss-spellings
  – Missing data
  – Outdated data
  – data not obeying business rules
Traditional methods

- Integrity Constraints of database (ICs)
  1) Functional dependencies
    - Example 1: Primary keys, foreign keys
    - Example 2: $[\text{CountryCode, AreaCode}] \rightarrow [\text{City}]$
  2) Conditional functional dependencies
    - Example: $[\text{CountryCode}=44, \text{AreaCode}] \rightarrow [\text{City}]$
      - Many others
Traditional methods

- Integrity Constraints only *detect* errors – cannot fix them!
  - Domain experts needed to find repair
    - → costly, manual

- Many error-detection algorithms have high time-complexity

  Example
  - Detecting duplicates is combinatorial problem

- NOSQL databases usually don't use ICs!
Traditional methods

● Problem of detecting errors
  – Scalability
    ● BigDansing addresses the *scalability* problem

● Problem of correcting errors
  – Accuracy
    ● KATARA addresses the *accuracy* problem
Instead of domain experts

- Knowledge Bases:
  - Example: Yago, DBPedia (Wikipedia)

- Crowdsourcing
KATARA

Overview of workflow
KATARA

Contributions

- Table pattern definition and discovery
  - new rank-join based algorithm to efficiently discover table patterns with high scores
- Table pattern validation via crowdsourcing
  - entropy-based scheduling algorithm
- Data annotation
  - algorithm to generate top-k possible repairs for those erroneous data
KATARA

Example: Pattern discovery and annotation
KATARA

Example: question for the crowd

Q1: What is the most accurate type of the highlighted column?
(A, B, C, D, E, F, ...)
(Rossi, Italy, Rome, Verona, Italian, Proto, ...)
(Pirlo, Italy, Madrid, Juve, Italian, Flero, ...)

- country
- state
- economy
- none of the above
KATARA

- Experimental study results
  - Pattern discovery

<table>
<thead>
<tr>
<th></th>
<th>Support P</th>
<th>Support R</th>
<th>MaxLike P</th>
<th>MaxLike R</th>
<th>PGM P</th>
<th>PGM R</th>
<th>RankJoin P</th>
<th>RankJoin R</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiTables</td>
<td>.54</td>
<td>.59</td>
<td>.62</td>
<td>.68</td>
<td>.60</td>
<td>.67</td>
<td>.78</td>
<td>.86</td>
</tr>
<tr>
<td>WebTables</td>
<td>.65</td>
<td>.64</td>
<td>.63</td>
<td>.62</td>
<td>.77</td>
<td>.77</td>
<td>.86</td>
<td>.84</td>
</tr>
<tr>
<td>RelationalTables</td>
<td>.51</td>
<td>.51</td>
<td>.71</td>
<td>.71</td>
<td>.53</td>
<td>.53</td>
<td>.77</td>
<td>.77</td>
</tr>
<tr>
<td><strong>Yago</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiTables</td>
<td>.56</td>
<td>.70</td>
<td>.71</td>
<td>.89</td>
<td>.61</td>
<td>.77</td>
<td>.71</td>
<td>.89</td>
</tr>
<tr>
<td>WebTables</td>
<td>.65</td>
<td>.69</td>
<td>.80</td>
<td>.84</td>
<td>.76</td>
<td>.80</td>
<td>.82</td>
<td>.87</td>
</tr>
<tr>
<td>RelationalTables</td>
<td>.64</td>
<td>.67</td>
<td>.81</td>
<td>.86</td>
<td>.74</td>
<td>.77</td>
<td>.81</td>
<td>.86</td>
</tr>
<tr>
<td><strong>DBPedia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Pattern discovery precision and recall
KATARA

- Experimental study results
  - Effectiveness of possible repairs

<table>
<thead>
<tr>
<th></th>
<th>KATARA (Yago)</th>
<th>KATARA (DBPedia)</th>
<th>EQ</th>
<th>SCARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>1.0</td>
<td>1.0</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td>Soccer</td>
<td>N.A.</td>
<td>0.97</td>
<td>0.29</td>
<td>0.66</td>
</tr>
<tr>
<td>University</td>
<td>0.95</td>
<td>1.0</td>
<td>0.18</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 6: Data repairing precision and recall (Relational Tables)

<table>
<thead>
<tr>
<th></th>
<th>KATARA (Yago)</th>
<th>KATARA (DBPedia)</th>
<th>EQ</th>
<th>SCARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiTables</td>
<td>1.0</td>
<td>1.0</td>
<td>0.30</td>
<td>N.A.</td>
</tr>
<tr>
<td>WebTables</td>
<td>1.0</td>
<td>1.0</td>
<td>0.46</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

Table 7: Data repairing precision and recall (WikiTables and WebTables)
KATARA

- Experimental study results
  - Pattern validation

Figure 7: Pattern validation P/R (WebTables)
Discussion

● Strengths:
  – Automatic finding of repairs
  – Easy-to-answer questions presented to crowd

● Weaknesses:
  – If data is special-purpose, no KBs available
  – Not any crowd can be used!
Summary

- Dirty data is a big problem
- Even bigger problem with Big Data
- Traditional methods such as ICs don't scale to Big Data
- Knowledge Bases and crowdsourcing can help finding accurate repairs