Data X-Ray: A diagnostic tool for data errors

Xiaolan Wang
Xin Luna Dong
Alexandra Meliou
MANY APPLICATIONS RELY ON DATA

Data is not perfect! Erroneous data can be extremely costly!

Knowledge graph (www.google.com)

Social network analytics

Shopping systems of retail companies
KNOWLEDGE VAULT [Dong14]

3.0 billion extracted triples
More than 70% are wrong

prKB

Extraction System

Fusion

Extractor
Extractor
Extractor

Web Sources

TXT
DOM
TBL
ANO
KNOWLEDGE VAULT [Dong14]

Traditional method: identify errors and drop them

Extraction System

Fusion

Extractor

Extractor

Extractor

... ...

Extractor

Web Sources

TXT DOM TBL ANO

prKB Perfect KB
**Errors are Systematic**

- **Extractor**
  - Bad extraction rules

- **Fusion**

- **prKB** → **Perfect KB**

- **Web Sources**
  - TXT, DOM, TBL, ANO

- Faulty information

---

**KNOWLEDGE VAULT [Dong14]**
KNOWLEDGE VAULT [Dong14]

Continue to generate erroneous data

Extraction System

Fusion

Extractor

Extractor

Extractor

Extractor

Web Sources

TXT

DOM

TBL

ANO

prKB

prKB

prKB

prKB
KNOWLEDGE VAULT [Dong14]

Diagnose root reason for errors

Extraction System

Extractor

Fusion

Extractor

Web Sources

TX

ANO
REAL-WORLD SYSTEMATIC ERRORS

Context: Date of birth of athletes extracted from besoccer.com is set to default value 1986_02_18

(besoccer.com, date_of_birth, 1986_02_18)

# Triples 630
Error Rate 100%

Default Value Error
REAL-WORLD SYSTEMATIC ERRORS

(Extractor S, obj: Baseball Coach)
# Triples 674,000
Error Rate 89.3%
Context: reconciling all coaches to baseball coaches
E.g., [Bob Barton, profession, Baseball Coach]
REAL-WORLD SYSTEMATIC ERRORS

(Extractor T, pred:namesakes, obj:the county)
# Triples 4878
Error Rate 99.8%
E.g., [Salmon P. Chase, namesakes, the county]
Contexts: The county was named for Salmon P. Chase, former senator and governor of Ohio
Q: Can we treat the error triples as a diagnosis?  
A: No; for two reasons:  
• Too many erroneous triples (more than 2B in KV)  
• Due to a variety of errors

<table>
<thead>
<tr>
<th>Knowledge triple</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Domenico Modugno, DoB, 01/09/1958&gt;</td>
<td>False</td>
</tr>
<tr>
<td>&lt;Bert Kaempfert, DoB, 09/01/1961&gt;</td>
<td>False</td>
</tr>
<tr>
<td>&lt;The Singing Nun, DoB, 07/12/1963&gt;</td>
<td>False</td>
</tr>
<tr>
<td>&lt;Paul Mauriat, DoB, 10/02/1963&gt;</td>
<td>False</td>
</tr>
<tr>
<td>&lt;Shocking Blue, DoB, 02/07/1968&gt;</td>
<td>True</td>
</tr>
<tr>
<td>&lt;U2, DoB, 05/16/1987&gt;</td>
<td>True</td>
</tr>
</tbody>
</table>

Leveraging on existing data cleaning methods [Abiteboul99, Fan08, Kalashnikov06, Rahm00, Raman01]
## WHAT IS A DIAGNOSIS?

<table>
<thead>
<tr>
<th>Knowledge triple</th>
<th>Correct?</th>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
<th>Web source</th>
<th>Extractor</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Domenico Modugno, DoB, 01/09/1958&gt;</td>
<td>False</td>
<td>People/D.M.</td>
<td>Bio/DoB</td>
<td>Date/01091958</td>
<td>euromusicxx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Bert Kaempfert, DoB, 09/01/1961&gt;</td>
<td>False</td>
<td>People/B.K.</td>
<td>Bio/DoB</td>
<td>Date/09011961</td>
<td>euromusicxx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;The Singing Nun, DoB, 07/12/1963&gt;</td>
<td>False</td>
<td>People/TSN</td>
<td>Bio/DoB</td>
<td>Date/07121963</td>
<td>euromusicxx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Paul Mauriat, DoB, 10/02/1963&gt;</td>
<td>False</td>
<td>People/P.M.</td>
<td>Bio/DoB</td>
<td>Date/10021963</td>
<td>euromusicxx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Shocking Blue, DoB, 02/07/1968&gt;</td>
<td>True</td>
<td>People/S.B.</td>
<td>Bio/DoB</td>
<td>Date/02071968</td>
<td>wiki.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;U2, DoB, 05/16/1987&gt;</td>
<td>True</td>
<td>People/U2</td>
<td>Bio/DoB</td>
<td>Date/05161987</td>
<td>wiki.com</td>
<td>Extractor 1</td>
</tr>
</tbody>
</table>

**Group error data:**

Date from website (euromusicxx.com) extracted by Extractor 1 is wrong. (Bad extraction rule: use U.S. date format rule to extract date information from European website).
## WHAT IS A DIAGNOSIS?

### Knowledge triple

<table>
<thead>
<tr>
<th>Knowledge triple</th>
<th>Correct?</th>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
<th>Web source</th>
<th>Extractor</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Domenico Modugno, DoB, 01/09/1958&gt;</td>
<td>False</td>
<td>People/ D.M.</td>
<td>Bio/DoB</td>
<td>Date/ 01091958</td>
<td>euromusic xx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Bert Kaempfert, DoB, 09/01/1961&gt;</td>
<td>False</td>
<td>People/ B.K.</td>
<td>Bio/DoB</td>
<td>Date/ 09011961</td>
<td>euromusic xx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;The Singing Nun, DoB, 07/12/1963&gt;</td>
<td>False</td>
<td>People/ TSN</td>
<td>Bio/DoB</td>
<td>Date/ 07121963</td>
<td>euromusic xx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Paul Mauriat, DoB, 10/02/1963&gt;</td>
<td>False</td>
<td>People/ P.M.</td>
<td>Bio/DoB</td>
<td>Date/ 10021963</td>
<td>euromusic xx.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;Shocking Blue, DoB, 02/07/1968&gt;</td>
<td>True</td>
<td>People/ S.B.</td>
<td>Bio/DoB</td>
<td>Date/ 02071968</td>
<td>wiki.com</td>
<td>Extractor 1</td>
</tr>
<tr>
<td>&lt;U2, DoB, 05/16/1987&gt;</td>
<td>True</td>
<td>People/U2</td>
<td>Bio/DoB</td>
<td>Date/ 05161987</td>
<td>wiki.com</td>
<td>Extractor 1</td>
</tr>
</tbody>
</table>

### Input1: Element And its correctness

### Input2: Features Combination of meta-data information

### Output (diagnosis): set of features

Which diagnosis is the best?
**DATA XRAY: COST MODEL**

**Bayesian estimate of causal likelihood**

\[
Pr(F|E) = \prod_{f_i \in F} \alpha \epsilon_i \left(1 - \epsilon_i \right)
\]

**Cost Model:**

- **Conciseness:** fewer features preferred
- **Specificity:** higher error rate preferred
- **Consistency:** fewer true elements preferred

**Theorem 1:** Derive a diagnosis with minimum cost is NP-Complete
**DATAXRAY: ALGORITHM**

- Top-down iterative traversal

**Theorem 2:** The DataXRay traversal has linear complexity in the number of features; with $O(\# \text{ of features})$ approximation.
EVALUATION (ReVerb ClueWeb Extraction dataset)

DataXRay vs. SetCover [Chvatal79]

Execution time: 0.43 sec vs. 3 sec
EVALUATION (ReVerb ClueWeb Extraction dataset)

DataXRay vs. RedBlue [Peleg07]

Execution time: 0.43 sec vs. 4.2 sec

DataXRay vs. RedBlue [Peleg07]

Recall Precision F-measure
0.0 0.0 0.0
0.2 0.2 0.2
0.4 0.4 0.4
0.6 0.6 0.6
0.8 0.8 0.8
1.0 1.0 1.0

DataXRay+Greedy Greedy RedBlue DataAuditor FeatureSelection

Finer-granularity features preferred
EVALUATION (ReVerb ClueWeb Extraction dataset)

DataXRay vs. FeatureSelection [Tibshirani96, Ng04]

Execution time: 0.43 sec vs. 5.5 sec

Target on predication
Redundant features
Low error rate features
EVALUATION SUMMARY

- DataXRay is effective several real-world scenarios
  - Extraction errors, traffic incidents, ...

- DataXRay is better than alternative algorithms
  - Classification, summarization, set cover methods

- DataXRay is robust under different parameters and settings
  - Different error rate, feature failure, ...

- DataXRay is parallelizable in MapReduce
Takeaways

- Diagnosis is different than cleaning
  - Reason about **root cause** of data errors.

- Defined a good diagnosis
  - Cost function based on Bayesian analysis: **Conciseness, Specificity, Consistency.**

- Designed a scalable algorithm
  - Leverage the feature hierarchy.
  - The **top-down iterative algorithm** is efficient and easy to parallelize.
References

References

Reference
