FINDING QUALITY IN QUANTITY: THE CHALLENGE OF DISCOVERING VALUABLE SOURCES FOR INTEGRATION

Theodoros Rekatsinas
University of Maryland

Amol Deshpande, Xin Luna Dong, Lise Getoor and Divesh Srivastava
DATA, DATA, DATA …

Clean
Integrate
Analyze

publicly available data
structured in a way
that enables the data
to be fully
discoverable
and usable by
end users
DATA, DATA, DATA …

Clean
Analyze
Integrate

Knowledge Bases

Outbreak Prediction

Business Analysis

Stock Price Prediction
IN REALITY …

Clean

Analyze

Integrate
IN REALITY …

Clean

Integrate

Analyze
Cleaning and integrating data takes time and costs money!

Things only become worse when using data from low quality sources!
A REAL EXAMPLE

Knowledge-base construction in Google

State-of-the-art automatic knowledge extraction from Web
accu=30% [KV KDD`14/Sonya VLDB`14]

State-of-the-art fusion on top
prec=90%, recall=20% [KV KDD`14/Sonya VLDB`14]

Human curation to increase accuracy and coverage

Select sources carefully to focus resources!
INFLUENCING FACTORS

Data

Context
LOW QUALITY SOURCES

Data

Low coverage
High delays - staleness
Erroneous information

Biased information

Polarity:
-1 negative
0 neutral
1 positive

Subjectivity:
0 objective
1 subjective

Context
CONTEXT MATTERS

Fantasy politics season coming to a close

By Zach McGann | ESPN.com

Barack Obama, a noted sports fan, is a surefire first-round pick in MTV’s "Fantasy Election"
WE ARE IN NEED OF...

Data Source Management Systems

- Index the content of sources
- Build quality profiles
WE ARE IN NEED OF…

Data Source Management Systems

- Find relevant sources to user queries.
- Find sources that if combined, maximize the quality of integrated data.
- Explore different solutions.
REASONING ABOUT CONTENT

Data sources have diverse data domains.

Users interested in different data domains.

Use a knowledge base (KB) as back-end to reason about the content of sources and user queries.
Extend KB with a Correspondence Graph.

Context Clusters group instances and concepts.

Detect c-clusters using latent variable learning or frequent itemset mining.
Build source quality profiles per context cluster.

Compare source content with integrated content of all relevant sources.
SOURCE SIGHT

A data source management system for news stories (events).

News articles extracted from EventRegistry.com and originate from news papers, blogs, and social media.

Content semantically annotated using OpenCalais by Thomson Reuters.
RANKING IS NOT ENOUGH…

Entities: Obama, Topic: War_Conflict

<table>
<thead>
<tr>
<th>Source Ranking</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>nypost.com</td>
<td>0.42</td>
</tr>
<tr>
<td>nymag.com</td>
<td>0.37</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>0.37</td>
</tr>
<tr>
<td>csmonitor.com</td>
<td>0.32</td>
</tr>
<tr>
<td>cleveland.com</td>
<td>0.28</td>
</tr>
<tr>
<td>washingtonexaminer.com</td>
<td>0.23</td>
</tr>
<tr>
<td>gawker.com</td>
<td>0.20</td>
</tr>
<tr>
<td>democracynow.org</td>
<td>0.17</td>
</tr>
<tr>
<td>blogtown.portlandmercury.com</td>
<td>0.11</td>
</tr>
<tr>
<td>nydailynews.com</td>
<td>0.11</td>
</tr>
</tbody>
</table>
RANKING IS NOT ENOUGH...

Entities: Obama, Topic: War_Conflict

Combining Sources

nypost.com (ranked 1st),
nymag.com (ranked 2nd)
Coverage: 0.48

nypost.com (ranked 1st),
business-standard.com (not in top-10)
Coverage: 0.52
REASON ABOUT SETS

Perform source selection [DSS VLDB`13, RDS SIGMOD`1...]

Find the set of sources that maximizes the quality of integrated data while minimizing the overall cost.

But there are multiple quality metrics.

Coverage, Timeliness, Bias, Accuracy

How can we reason about different metrics?
PARETO OPTIMALITY

Source selection as multi-variate optimization.

Goal: find pareto optimal sets of sources
PARETO OPTIMALITY

Source selection as multi-variate optimization.

Goal: find pareto optimal sets of sources

Finding the pareto front is hard!
SOURCE SIGHT DEMO

SOURCE SIGHT

DISCOVERING VALUABLE SOURCES FOR INTEGRATION

LEARN MORE
CHALLENGES

The content and quality of data sources changes over time. How can we update the content and quality profiles efficiently?

How can we build quality profiles (e.g., via sampling) that come with rigorous guarantees?

How can we provide succinct descriptions of the source characteristics?

How can we provide users with explanations? Why does this source appear in my result?
CONCLUSIONS

Reasoning about the quality of data sources and their relevance to user queries is crucial.

Data source management systems should support diverse integrations tasks and allow users to understand the quality of integrated data.

We presented Source Sight a prototype data source management system.

Thank you!
thodrek@cs.umd.edu