HOLISTIC DATA INTEGRATION FOR BIG DATA

ERHARD RAHM,
UNIVERSITY OF LEIPZIG,
AUGUST 2016

www.scads.de
Two Centers of Excellence for Big Data in Germany

- ScaDS Dresden/Leipzig
- Berlin Big Data Center (BBDC)

ScaDS Dresden/Leipzig:
Competence Center for Scalable Data Services and Solutions Dresden/Leipzig

- scientific coordinators: Nagel (TUD), Rahm (UL)
- start: Oct. 2014
- duration: 4 years (option for 3 more years)
STRUCTURE OF THE CENTER

Life sciences
Material and Engineering sciences
Environmental / Geo sciences
Digital Humanities
Business Data

Big Data Life Cycle Management and Workflows

Data Quality / Data Integration
Knowledge Extraktion
Visual Analytics

Efficient Big Data Architectures
DATA QUALITY AND INTEGRATION

- Parallel execution of comprehensive data integration workflows
- Learning based configuration of integration workflows

Continuous changes in thousands of data sources

“This tool (Dedoop) by far shows the most mature use of MapReduce for data deduplication”

www.hadoopsphere.com
LARGE SCALE GRAPH ANALYTICS

- GRADOOP framework
  - End-to-end graph data management and analytics
  - Extended property graph data model (EPGM)
  - Powerful graph operators for data integration and graph analytics
  - Analytics language GRALA
  - Parallel execution on Hadoop clusters
  - Open-source: www.gradoop.org
AGENDA

- Introduction
  - Big Data
  - Scalable data integration

- Holistic data integration: use cases
- Holistic entity resolution for Linked Data
- Summary
BIG DATA CHALLENGES

Volume
Petabytes / exabytes of data

Velocity
fast analysis of data streams

Variety
heterogeneous data of different kinds

Veracity
high data quality

Value
useful analysis results
BIG DATA ANALYSIS PIPELINE

Data acquisition → Data extraction/cleaning → Data integration/annotation → Data analysis and visualization → Interpretation

Heterogeneity • Volume • Velocity • Privacy • Human collaboration
2 LEVELS OF DATA INTEGRATION

- **Metadata (schema/ontology) level**
  - **Schema Matching**: find correspondences between source schemas and target schema
  - **Schema Merge**: combine source schemas into integrated target schema

- **Instance (entity) level**
  - transform heterogeneous source data into uniform representation
  - identify and resolve data quality problems
  - identify and resolve equivalent instance records: **object matching / entity resolution / link discovery**
  - Fusion of matching objects
Identification of semantically equivalent objects

- within one data source or between different sources

Original focus on structured (relational) data, e.g. customer data

<table>
<thead>
<tr>
<th>Cno</th>
<th>LastName</th>
<th>FirstName</th>
<th>Gender</th>
<th>Address</th>
<th>Phone/Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Smith</td>
<td>Christoph</td>
<td>M</td>
<td>23 Harley St, Chicago IL, 60633-2394</td>
<td>333-222-6542 / 333-222-6599</td>
</tr>
<tr>
<td>493</td>
<td>Smith</td>
<td>Kris L.</td>
<td>F</td>
<td>2 Hurley Place, South Fork MN, 48503-5998</td>
<td>444-555-6666</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CID</th>
<th>Name</th>
<th>Street</th>
<th>City</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Kristen Smith</td>
<td>2 Hurley Pl</td>
<td>South Fork, MN 48503</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>Christian Smith</td>
<td>Hurley St 2</td>
<td>S Fork MN</td>
<td>1</td>
</tr>
</tbody>
</table>
DUPLICATE WEB ENTITIES: PUBLICATIONS

Data cleaning: Problems and current approaches
Cited by 1234 - Related articles - All 37 versions

Data cleaning: Problems and current approaches *
ERHH Do, E Rahm - IEEE Data Engineering Bulletin, 2000
Cited by 12 - Related articles

Hai Do H.: Data Cleaning: Problems and Current approaches *
E Rahm - Bulletin of the Technical Committee on Data ..., 2000
Cited by 7 - Related articles

Problems and Current Approaches *
D Cleaning - Erhard Rahm, Hong Hai Do. IEEE Data Engineering ..., 2000
Cited by 6 - Related articles

Hong Hai Do *
E Rahm - IEEE Bulletin of the Technical Committee on Data ..., 2000
Cited by 5 - Related articles

Do Hong Hai. Data Cleaning: Problems and Current Approaches *
E Rahm - IEEE Data Engineering Bulletin, 2000
Cited by 4 - Related articles

Data Cleaning: Problems & Current Approaches *
D Hang-Hai, R Erhard - IEEE bulletin of the technical committee on Data ..., 2000
Cited by 4 - Related articles

Data Cleaning: Problems and Current Approaches. IEEE Techn *
E Rahm, HH Do - Bulletin on Data Engineering, 2000
Cited by 3 - Related articles
GENERAL OBJECT MATCHING WORKFLOW

- **Blocking**
- **Similarity Computation**
- **Match Classification**

Matching object pairs

- **Special cases:**
  - only one input data source
  - 2 clean (duplicate-free) sources -> 1:1 mapping
  - 1 clean, 1 unclean source -> 1:n mapping
  - 2 unclean sources -> n:m mappings
• Large-scale matching
  • reduce search space, e.g. utilizing blocking techniques
  • massively parallel processing (Hadoop clusters, GPUs, etc.)

• Holistic data integration
  • support for many data sources, not only 1 or 2
  • binary integration approaches do not scale

• Data quality
  • unstructured, semi-structured sources
  • need for data cleaning and enrichment

• Privacy for sensitive data
  • privacy-preserving record linkage and data mining
PARALLEL OBJECT MATCHING WITH MAP/REDUCE

Map Phase: Blocking

Reduce Phase: Matching
- Parallel execution of data integration/match workflows with Hadoop
- Powerful library of match and blocking techniques
- Learning-based configuration
- GUI-based workflow specification
- Automatic generation and execution of Map/Reduce jobs on different clusters
- Automatic load balancing for optimal scalability
- Iterative computation of transitive closure

“This tool by far shows the most mature use of MapReduce for data deduplication”
www.hadoopsphere.com
Introduction

Holistic data integration: use cases
- Linked Open Data (LOD)
- Holistic schema matching / ontology integration
- Web tables
- Matching of product offers
- Knowledge graphs
- Comparison

Holistic entity resolution for Linked Data

Summary
Scalable approaches for integrating N data sources (N >> 2)

Increasing need due to numerous sources, e.g., from the web
- hundreds of LOD sources (Data Web)
- many thousands of web shops
- many millions of web tables

Large open data /metadata/mapping repositories

Pairwise matching does not scale
- 200 sources -> 20,000 mappings
- many linked data sources containing ontologies and associated instances
- `sameAs` Links (mappings) between ontology concepts and instances for data integration
- federated SPARQL queries to traverse links and access data from different sources
Considered LD tools (sorted by year of initial publication)

<table>
<thead>
<tr>
<th>System / initial publication</th>
<th>Year</th>
<th>Institution</th>
<th>Learning-based</th>
<th>OAEI IM participation</th>
<th>Support for pure ontology matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiMOM [66]</td>
<td>2004</td>
<td>Univ. of Tsinghua, China</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>KnoFuss [46]</td>
<td>2007</td>
<td>Open Univ. Milton Keynes, UK</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>AgreementMaker [9]</td>
<td>2009</td>
<td>Univ. of Illinois at Chicago, USA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Silk [69]</td>
<td>2009</td>
<td>FU Berlin, Germany</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CODI [44]</td>
<td>2010</td>
<td>Univ. of Mannheim, Germany</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LIMES [39]</td>
<td>2011</td>
<td>Univ. of Leipzig, Germany</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERIMI [3]</td>
<td>2011</td>
<td>Delft Univ. of Techn., Netherlands</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhishi.links [48]</td>
<td>2011</td>
<td>Shanghai Jiao Tong Univ., China</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SLINT+ [43]</td>
<td>2012</td>
<td>Nat. Inst. of Informatics, Japan</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RuleMiner [47]</td>
<td>2012</td>
<td>Shanghai Jiao Tong Univ., China</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Characteristics of proposed LD frameworks

<table>
<thead>
<tr>
<th></th>
<th>RMOM</th>
<th>AgreementMaker</th>
<th>CODI</th>
<th>LogMap</th>
<th>SERIMI</th>
<th>Zhishi links</th>
<th>SLINT+</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Input</strong></td>
<td>RDF, OWL</td>
<td>SBA RQL</td>
<td>RDF, OWL</td>
<td>RDF, OWL</td>
<td>SPARQL</td>
<td>RDF</td>
<td>RDF</td>
</tr>
<tr>
<td><strong>Supported link types</strong></td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
<td>owl:sameAs</td>
</tr>
<tr>
<td><strong>Configuration</strong></td>
<td>adaptive</td>
<td>manual</td>
<td>manual</td>
<td>manual</td>
<td>adaptive</td>
<td>manual</td>
<td>adaptive</td>
</tr>
<tr>
<td>- matcher combination</td>
<td>weighted</td>
<td>weighted</td>
<td>weighted</td>
<td>weighted</td>
<td>weighted</td>
<td>weighted</td>
<td>weighted</td>
</tr>
<tr>
<td>- average combination</td>
<td>average</td>
<td>average</td>
<td>average</td>
<td>average</td>
<td>average</td>
<td>average</td>
<td>average</td>
</tr>
<tr>
<td><strong>Runtime optimization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Blocking</td>
<td>indexing</td>
<td>indexing</td>
<td></td>
<td>indexing</td>
<td>indexing</td>
<td>indexing</td>
<td>indexing</td>
</tr>
<tr>
<td>- Filtering</td>
<td>indexing</td>
<td>indexing</td>
<td></td>
<td>indexing</td>
<td>indexing</td>
<td>indexing</td>
<td>indexing</td>
</tr>
<tr>
<td><strong>String similarity measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Further similarity measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Structure matcher</strong></td>
<td></td>
<td>semantic similarity</td>
<td>iterative anchor-based mapping generation</td>
<td>iterative anchor-based mapping generation</td>
<td></td>
<td>semantic similarity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Use of</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- external dictionaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- existing mappings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Post-processing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Parallel processing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GUI/web interface/API</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Download Tool/Source</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Open Source</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Learning-Based Link Discovery Tools

Characteristics of learning-based LD frameworks. "-" means not existing, "*" investigated in [20], but not available in current release.

<table>
<thead>
<tr>
<th>Data Input</th>
<th>KnoFuss</th>
<th>Silk</th>
<th>LIMES</th>
<th>RuleMiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Input</td>
<td>RDF, SPARQL</td>
<td>RDF, SPARQL, CSV</td>
<td>RDF, SPARQL, CSV</td>
<td>RDF</td>
</tr>
<tr>
<td>Supported linktypes</td>
<td>owl:sameAs</td>
<td>owl:sameAs, user-specified others</td>
<td>owl:sameAs, user-specified others</td>
<td>owl:sameAs</td>
</tr>
<tr>
<td>Configuration</td>
<td>manual (match rules), unsupervised learning (genetic programming)</td>
<td>manual (match rules), supervised learning (genetic programming, active learning)</td>
<td>manual (match rules), supervised learning (genetic programming, active learning), unsupervised (genetic programming)</td>
<td>adaptive (match rules), supervised learning (expectation maximization)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Runtime optimization</th>
<th>- Blocking</th>
<th>multi-dimensional space tiling</th>
<th>- indexing</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Filtering</td>
<td>indexing</td>
<td>-</td>
<td>indexing</td>
</tr>
<tr>
<td>String similarity measures</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Further similarity measures</td>
<td>-</td>
<td>numeric, date equality</td>
<td>geographical coordinates, numeric, date equality</td>
</tr>
<tr>
<td>Structure matcher</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Use of</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>- external dictionaries</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>- existing mappings</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Post-processing</td>
<td>one-to-one mapping</td>
<td>-</td>
<td>Stable marriage, hospital-resident</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parallel Processing</th>
<th>-</th>
<th>MapReduce (MapReduce)*</th>
<th>MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUI/web interface/API</td>
<td>-/-/-</td>
<td>✓ ✓ ✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Download Tool/Source</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ -</td>
</tr>
<tr>
<td>Open Source project</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
LD TOOLS: OBSERVATIONS

- mostly simple (attribute/property-based) matchers
  - limited use of context and existing links / synonym information
- limited use of blocking
  - use of filtering/indexing to speed up string comparisons
- mostly support for semi-automatic configurations, e.g. utilizing learning-based techniques
- general availability of most tools
- need for more comparative evaluations, especially regarding large-scale datasets
- binary linking only: no support for holistic data integration
- creation of a mediated schema
  - holistic matching between N simple schemas, e.g., web forms
  - virtual data integration: meta-search
- holistic matching based on *clustering* of similar attributes
  - utilize high name similarity between schemas
  - similar names within a schema are mismatches (e.g. first name, last name)
- Development of an integrated domain ontology (e.g., UMLS)
  - physical metadata integration
  - tens of source ontologies
  - clustering of synonymous concepts (synsets)
  - largely manual integration effort

source: Fiszman, 2012 “Natural Language Processing and the National Library of Medicine”
- huge data collections with (mostly unrelated) up to millions of files or tables from numerous domains
Web contains hundreds of millions tables
- only 1% relational tables + vertical tables about one entity*
- several corpora with huge number of heterogenous tables

Longest rivers

<table>
<thead>
<tr>
<th></th>
<th>River</th>
<th>Length (km)</th>
<th>Length (miles)</th>
<th>Drainage area (km²)[citation needed]</th>
<th>Average discharge (m³/s)[citation needed]</th>
<th>Outflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nile – Kagera[a 1]</td>
<td>6,853</td>
<td>4,258</td>
<td>3,254,555</td>
<td>5,100</td>
<td>Mediterranean</td>
</tr>
<tr>
<td>2</td>
<td>Amazon – Ucayali – Apurimac[a 1]</td>
<td>6,902</td>
<td>4,345</td>
<td>7,050,000</td>
<td>210,000</td>
<td>Atlantic Ocean</td>
</tr>
<tr>
<td>3</td>
<td>Yangtze (Chang Jiang)</td>
<td>6,300</td>
<td>3,917</td>
<td>1,800,000</td>
<td>31,900</td>
<td>East China Sea</td>
</tr>
<tr>
<td>4</td>
<td>Mississippi–Missouri–Jefferson</td>
<td>6,275</td>
<td>3,902</td>
<td>2,980,000</td>
<td>16,200</td>
<td>Gulf of Mexico</td>
</tr>
<tr>
<td>5</td>
<td>Yenisei–Angara–Solenge</td>
<td>5,539</td>
<td>3,445</td>
<td>2,580,000</td>
<td>19,600</td>
<td>Kara Sea</td>
</tr>
</tbody>
</table>

- **Query support**
  - find related tables for keywords
  - augment tables by additional attributes

- **Need to add semantics**
  - attributes need to be annotated, e.g., with knowledge graph
  - table contents described in surrounding text
  - Vertical tables: identify key column vs. property column

- **Table augmentation**: find coherent attributes from other tables that can extend a given table

<table>
<thead>
<tr>
<th>Company</th>
<th>Revenue 2012</th>
<th>Revenue 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank of China</td>
<td>x1</td>
<td>x2</td>
</tr>
<tr>
<td>Deutsche Bank</td>
<td>y1</td>
<td>y2</td>
</tr>
<tr>
<td>Banco do Brasil</td>
<td>z1</td>
<td>z3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Telco companies</th>
<th>Revenue 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>China Mobile</td>
<td>x</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>y</td>
</tr>
</tbody>
</table>

- many open challenges to better utilize table/dataset corpora
  - more comprehensive domain categorization and attribute annotation
  - more sophisticated matching of attributes based on instances + metadata
  - clustering and fusing related tables/datasets
  - derivation of domain-specific knowledge graphs
HOLISTIC INTEGRATION OF ENTITIES

- Entity search engines
  - clustering of matching entities (publications, product offers)
  - physical data integration
  - thousands of data sources

- Comparison / booking portals
  - clustered offers within (integrated) taxonomy
  - physical or virtual data integration
HOLISTIC DATA INTEGRATION USE CASE:
INTEGRATION OF PRODUCT OFFERS IN COMPARISON PORTAL

- thousands of data sources (shops/merchants)
- millions of products and product offers
- continuous changes
- many similar, but different products
- low data quality

Canon VIXIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x optical zoom
Flash card: SD HS1, 1x warranty, PAL 0.3 m
The VIXIA HF S10 delivers brilliant video and photos through a Canon exclusive 8.59 megapixel CMOS image sensor and the latest version of Canon's advanced image processor...

***** 12 reviews - Add to Shopping List

Canon VIXIA HF S10 IVIS Dual Flash Memory Camcorder
Canon HF S10 IVIS Dual Flash Memory Camcorder SPECIAL SALE PRICE: $899
Display both English/Japanese + we supply all English manuals in English as PDF. ...
Add to Shopping List

Canon VIXIA HF S10
Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video
Canon has a well-known and highly-regarded reputation for optical excellence, ...
Add to Shopping List

Canon VIXIA HF S100 Flash Memory Camcorder
Canon Video HF S100 Instant Rebate Receive $200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes $200) ...
Add to Shopping List

Canon VIXIA HF S10 Care & Cleaning
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning
Add to Shopping List
Input:

- new product offers
- existing product catalog with associated products and offers

1. preprocessing/ data cleaning:
   - extraction and consolidation of manufacturer info
   - extraction of product codes

2. (learning-based) categorization of product offers
   - determine product (entity) type

3. (learning-based) matching of product offers per product category

LEARNING-BASED MATCH APPROACH

1. Pre-processing
   - Product Code Extraction
   - Manufacturer Cleaning
   - Automatic Categorization

2. Training
   - Training Data Selection
   - Matcher Application
   - Classifier Learning

3. Application
   - Blocking (Manufacturer + Category)
   - Matcher Application
   - Classification

Product Offers

Classifier

Product Match Result
- web-scale integration and categorization of heterogeneous entities
Enhanced search results

University of Leipzig
www.uni-leipzig.de/ • Translate this page
Offizieller Intemetauftritt mit Vorstellung der Leipziger Universität mit umfangreichen Informationen zu Forschung und Lehre.

Results from uni-leipzig.de

Studiengänge
Kommunikations - Management
Science - Psychologie - Medizin

Fakultäten
14 Fakultäten der Universität.
Theologische Fakultät...

Bewerbung und Immatrikulierung
Sie sind hier: Studium • Bewerbung und ...

International Study
International students: Welcome. You are from abroad and you...

Leipzig University
University in Leipzig, Germany
Leipzig University, located in Leipzig in the Free State of Saxony, Germany, is one of the oldest universities in the world and the second-oldest university in Germany. Wikipedia

Address: Augustusplatz 10, 04109 Leipzig
Enrollment: 26,275 (2014)
Customer service: 0341 97108
Founded: December 2, 1409
President: Beate Schücking
Founders: William II, Margrave of Meissen, Frederick I, Elector of Saxony, Wilhelm Wundt

Profiles
LinkedIn

Notable alumni
Angela Merkel
Wolfgang von Goethe
Gottfried Wilhelm Leibniz
Richard Wagner
Gotthold Ephraim Lessing

University of Leipzig (UniLeipzig) • Twitter
combines knowledge from numerous sources
  - Freebase, Wikipedia, CIA World fact book, ...
  - 2012: > 570 million entities, > 18 billion facts/relationships
uniform representation and semantic categorization of entities of many domains (web-scale) or only one domain, one enterprise etc.

- examples: DBPedia, Yago, Wikidata, Google KG, MS Satori, Facebook, ...
- entities and their metadata often extracted from other resources such as Wikipedia, Wordnet etc. as well as from normal web pages, documents, web searches etc.
- comprehensive taxonomies to categorize entities and their details
- extreme entity heterogeneity (attributes + values) even within domains
- very challenging data integration problems

Knowledge Graphs provide valuable background knowledge for

- enhancing entities (based on prior entity linking)
- improving data integration (e.g., by utilizing additional information)
- improving search results ...
## Use Cases for Holistic Data Integration

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Data Integration Type</th>
<th>#Domains</th>
<th>#Sources</th>
<th>Clustering?</th>
<th>Degree of Automated Data Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-search</td>
<td>virtual, metadata</td>
<td>1</td>
<td>low-mid</td>
<td>attributes</td>
<td>medium</td>
</tr>
<tr>
<td>Open data, web tables</td>
<td>physical, primarily metadata</td>
<td>many</td>
<td>very high</td>
<td>(possible)</td>
<td>high, but limited integration</td>
</tr>
<tr>
<td>Integrated ontology</td>
<td>physical, metadata</td>
<td>1+</td>
<td>low-mid</td>
<td>concepts</td>
<td>low-medium</td>
</tr>
<tr>
<td>Entity search engines</td>
<td>physical, data (+ metadata)</td>
<td>1</td>
<td>very high</td>
<td>entities</td>
<td>high</td>
</tr>
<tr>
<td>Booking portals</td>
<td>physical, data + metadata</td>
<td>1+</td>
<td>high</td>
<td>entities</td>
<td>high</td>
</tr>
<tr>
<td>Knowledge graphs</td>
<td>physical, data + metadata</td>
<td>many</td>
<td>low-high</td>
<td>entities + concepts/attributes</td>
<td>medium-high</td>
</tr>
</tbody>
</table>
Most scalable approaches are based on
- Physical data integration
- Integration of instance data rather than metadata integration

Clustering instead of mappings
- cluster of k matching objects represents $k^2/2$ correspondences
- cluster size limited by #sources (for duplicate-free sources)
- simplified fusion of corresponding objects
- incremental data integration: additional sources/objects only need to be matched with clusters instead of all other sources

-> need for clustering-based concept matching and instance matching
Introduction

Holistic data integration: use cases

Holistic entity resolution for Linked Data
  - High-level clustering approach
  - Clustering for linked data

Summary
HOLISTIC ENTITY RESOLUTION

- requirements
  - scalability to many data sources and high data volumes
  - dynamic addition of new sources/entities (data streams)
  - support for many entity types
  - high match quality
  - little or no manual interaction
- binary match approaches not sufficient
clustering-based approaches
- represent matching entities from k sources in single cluster
- determine *cluster representative* for further processing/matching

incremental addition/clustering of sources, e.g., starting with the largest data source

distinguish clusters by entity type (product categories, geographical entities, etc.)
- determine entity type for new entities (preprocessing)

utilize blocking to restrict number of clusters to match with
CLUSTERING-BASED ENTITY RESOLUTION

Data sources $D_i$ with new entities of different types

Cluster set for entity type $T_1$

Cluster set for entity type $T_k$

...
characteristics

- entity clusters instead of binary sameAs links
- utilization of existing (possibly noisy) link mappings
- support for different entity types

found clusters represent many additional links and fused knowledge of improved quality
- can be used to build knowledge graphs

Preprocessing
- normalization of semantic entity types
- pruning 1-to-many links

Clustering
- initial clustering: connected components
- cluster decomposition: type-based, similarity-based
- iterative cluster merge

Output
Set of clusters \( C \)

Use case: cluster geographical entities from 4 sources
- GeoNames
- DBpedia
- NY Times
- existing sameAs links partially wrong
  - equal name, similar geographic coordinates
  - but different type (lake vs. city)
<table>
<thead>
<tr>
<th>id</th>
<th>label</th>
<th>source</th>
<th>type</th>
<th>latitude</th>
<th>longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Lake Louise (Canada)</td>
<td>NYTimes</td>
<td>-</td>
<td>51.42</td>
<td>-116.23</td>
</tr>
<tr>
<td>1</td>
<td>Lake Louise</td>
<td>GeoNames</td>
<td>gn:H.LK</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
<tr>
<td>2</td>
<td>Lake Louise, Alberta</td>
<td>DBpedia</td>
<td>db:Settlement</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Lake Louise alberta</td>
<td>FreeBase</td>
<td>fb:location.citytown</td>
<td>51.43</td>
<td>-116.16</td>
</tr>
<tr>
<td>4</td>
<td>Lake Louise (Alberta)</td>
<td>DBpedia</td>
<td>db:BodyOfWater</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
<tr>
<td>5</td>
<td>lake louise</td>
<td>FreeBase</td>
<td>fb:geography.lake</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
<tr>
<td>6</td>
<td>Mystic (Conn)</td>
<td>NYTimes</td>
<td>-</td>
<td>41.35</td>
<td>-71.97</td>
</tr>
<tr>
<td>7</td>
<td>N17632379615920</td>
<td>FreeBase</td>
<td>-</td>
<td>41.35</td>
<td>-71.97</td>
</tr>
<tr>
<td>8</td>
<td>Mystic</td>
<td>GeoNames</td>
<td>gn:P.PPL</td>
<td>41.35</td>
<td>-71.97</td>
</tr>
<tr>
<td>9</td>
<td>Black Hill</td>
<td>GeoNames</td>
<td>gn:T.HLL</td>
<td>53.54</td>
<td>-1.89</td>
</tr>
<tr>
<td>10</td>
<td>Black Hill</td>
<td>DBpedia</td>
<td>db:Mountain</td>
<td>53.53</td>
<td>-1.88</td>
</tr>
<tr>
<td>11</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>lgd:Peak</td>
<td>53.96</td>
<td>-1.85</td>
</tr>
<tr>
<td>12</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>lgd:Peak</td>
<td>54.69</td>
<td>-2.15</td>
</tr>
<tr>
<td>13</td>
<td>Kathmandu (nepal)</td>
<td>NYTimes</td>
<td>-</td>
<td>27.72</td>
<td>85.32</td>
</tr>
<tr>
<td>14</td>
<td>Kathmandu</td>
<td>GeoNames</td>
<td>gn:P.PPLC</td>
<td>27.7</td>
<td>85.32</td>
</tr>
<tr>
<td>15</td>
<td>Kathmandu</td>
<td>DBpedia</td>
<td>db:Settlement</td>
<td>27.7</td>
<td>85.33</td>
</tr>
<tr>
<td>16</td>
<td>Kathmandu</td>
<td>FreeBase</td>
<td>fb:location.citytown</td>
<td>27.7</td>
<td>85.37</td>
</tr>
</tbody>
</table>
Each source uses different entity types or no types
- e.g., Settlement, AdministrativeRegion, Park, Country, Island

Type harmonization using background knowledge

Remove links for incompatible types
- BodyOfWater != Settlement

<table>
<thead>
<tr>
<th>id</th>
<th>label</th>
<th>source</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Lake Louise (Canada)</td>
<td>NYTues</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>Lake Louise</td>
<td>GeoNames</td>
<td>BodyOfWater</td>
</tr>
<tr>
<td>2</td>
<td>Lake Louise, Alberta</td>
<td>DBpedia</td>
<td>Settlement</td>
</tr>
<tr>
<td>3</td>
<td>lake louise alberta</td>
<td>FreeBase</td>
<td>Settlement</td>
</tr>
<tr>
<td>4</td>
<td>Lake Louise (Alberta)</td>
<td>DBpedia</td>
<td>BodyOfWater</td>
</tr>
<tr>
<td>5</td>
<td>lake louise</td>
<td>FreeBase</td>
<td>BodyOfWater</td>
</tr>
<tr>
<td>6</td>
<td>Mystic (Conn)</td>
<td>NYTues</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>N17632379615920</td>
<td>FreeBase</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Mystic</td>
<td>GeoNames</td>
<td>Settlement</td>
</tr>
<tr>
<td>9</td>
<td>Black Hill</td>
<td>GeoNames</td>
<td>Mountain</td>
</tr>
<tr>
<td>10</td>
<td>Black Hill</td>
<td>DBpedia</td>
<td>Mountain</td>
</tr>
<tr>
<td>11</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>Mountain</td>
</tr>
<tr>
<td>12</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>Mountain</td>
</tr>
<tr>
<td>13</td>
<td>Katmandu (nepal)</td>
<td>NYTues</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>Kathmandu</td>
<td>GeoNames</td>
<td>Settlement</td>
</tr>
<tr>
<td>15</td>
<td>Kathmandu</td>
<td>DBpedia</td>
<td>Settlement</td>
</tr>
<tr>
<td>16</td>
<td>Kathmandu</td>
<td>FreeBase</td>
<td>Settlement</td>
</tr>
</tbody>
</table>
### Holistic Clustering (1): Check Duplicates

<table>
<thead>
<tr>
<th>ID</th>
<th>Location</th>
<th>Data Source</th>
<th>Category</th>
<th>Value1</th>
<th>Value2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Black Hill</td>
<td>GeoNames</td>
<td>Mountain</td>
<td>53.54</td>
<td>-1.89</td>
</tr>
<tr>
<td>10</td>
<td>Black Hill</td>
<td>DBpedia</td>
<td>Mountain</td>
<td>53.53</td>
<td>-1.88</td>
</tr>
<tr>
<td>11</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>Mountain</td>
<td>53.96</td>
<td>-1.85</td>
</tr>
<tr>
<td>12</td>
<td>Black Hill</td>
<td>LinkedGeoData</td>
<td>Mountain</td>
<td>54.69</td>
<td>-2.15</td>
</tr>
</tbody>
</table>
HOLISTIC CLUSTERING (2): CONN. COMPONENTS

Pre processing  Initial Clustering  Cluster Decomposition  Cluster Merge

Input

0 1 2 3
4 5
6 7 8
9 10 11 12
13 14 15 16

0 1 2 3
4 5
6 7 8
9 10 11
13 14 15 16

0 1 2 3
4 5
6 7 8
9 10 11
13 14 15 16
- Two split strategies for cluster refinement
  - Type-based grouping: ensure same entity type per cluster
  - Similarity-based refinement: eliminate entities with partially low intra-cluster similarity

- Determine cluster representatives with unified properties
• Connected components may contain entities of different types, e.g., via intermediate entities without type

• Create sub-clusters for each type

• Assign entities without type to cluster of best-matching entity

<table>
<thead>
<tr>
<th></th>
<th>Entity 1</th>
<th>Entity 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Lake Louise (Canada)</td>
<td>NYT times</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Lake Louise</td>
<td>GeoNames</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BodyOfWater</td>
</tr>
<tr>
<td>2</td>
<td>Lake Louise, Alberta</td>
<td>DBpedia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Settlement</td>
</tr>
<tr>
<td>3</td>
<td>lake louise alberta</td>
<td>FreeBase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Settlement</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>51.42, -116.23</td>
</tr>
<tr>
<td>1</td>
<td>51.41, -116.23</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>51.43, -116.16</td>
</tr>
</tbody>
</table>
- Iterative process to exclude low similarity entities from clusters
- Remove entity with lowest similarity from cluster if below threshold

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Mystic (Conn)    NYTimes    -    41.35    -71.97
7 N17632379615920  FreeBase   -    41.35    -71.97
8 Mystic           GeoNames   Settlement 41.35    -71.97
```
• **Iterative merging of similar clusters**
  
  • Similarity of cluster representatives above threshold
  
  • only consider pairs of clusters of the same type and with entities from different sources
### HOLISTIC CLUSTERING (4): SAMPLE CLUSTER MERGE

<table>
<thead>
<tr>
<th></th>
<th>Location</th>
<th>Source(s)</th>
<th>Type</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Lake Louise (Canada)</td>
<td>NYTimes</td>
<td>-</td>
<td>51.42</td>
<td>-116.23</td>
</tr>
<tr>
<td>1</td>
<td>Lake Louise</td>
<td>GeoNames</td>
<td>BodyOfWater</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
<tr>
<td>2</td>
<td>Lake Louise, Alberta</td>
<td>DBpedia</td>
<td>Settlement</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>lake louise albera</td>
<td>FreeBase</td>
<td>Settlement</td>
<td>51.43</td>
<td>-116.16</td>
</tr>
<tr>
<td>4</td>
<td>Lake Louise (Alberta)</td>
<td>DBpedia</td>
<td>BodyOfWater</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
<tr>
<td>5</td>
<td>lake louise</td>
<td>FreeBase</td>
<td>BodyOfWater</td>
<td>51.41</td>
<td>-116.23</td>
</tr>
</tbody>
</table>

**Node Descriptions:**

- **Node 0:**
  - Label: Lake Louise
  - Geo: 51.41, -116.23
  - Type: BodyOfWater
  - Sources: nyt, gn
  - Entities: 0, 1

- **Node 2:**
  - Label: lake louise alberta
  - Geo: 51.43, -116.16
  - Type: Settlement
  - Sources: dbp, fb
  - Entities: 2, 3

- **Node 4:**
  - Label: Lake Louise
  - Geo: 51.41, -116.23
  - Type: BodyOfWater
  - Sources: dbp, fb
  - Entities: 4, 5
AGENDA

- Introduction
- Holistic data integration: use cases
- Holistic entity resolution for Linked Data

- Summary
Big Data Integration
- Big Data poses new requirements for data integration (variety, volume, velocity, veracity)
- blocking and parallel matching for improved scalability, e.g. utilizing Hadoop-based approaches such as Dedoop

Holistic data integration
- combined integration of many sources (metadata + instances)
- clustering-based rather than mapping-based approaches
- utilization of corpora with datasets, schemas, and mappings
- construction of and linking to large knowledge graphs
- many research opportunities


REFERENCES (2)

- L. Kolb, A. Thor, E. Rahm: *Don’t Match Twice: Redundancy-free Similarity Computation with MapReduce*. Proc. 2nd Intl. Workshop on Data Analytics in the Cloud (DanaC), 2013