Entity-Centric Data Management

Philippe Cudré-Mauroux

eXascale Infolab, University of Fribourg
Switzerland
eXascale Infolab

• New lab @ U. of Fribourg, Switzerland
• Financed largely by the Swiss Federal State
• Extremely large-scale, non-relational data management (… mostly)
XI Projects

dipLODocus\[RDF\]

ZenCrowd

SciDB

dipLODocus\[RDF\]

ZenCrowd

SciDB

eXascale Infolab

DNS^3

eGovTeC

ScienceWISE
Entities
Information Management

• The story so far:
  – Strict separation between unstructured and structured data management infrastructures

![Diagram showing relationships between HTTP, Inverted Index, JDBC, SQL, and DBMS]
Information Integration

• Information integration is still the one of the biggest CS problem out there (according to many e.g., Gartner)

• Information integration typically requires some sort of mediation
  1. Unstructured Data: keywords, synsets
  2. Structured Data: global schema, transitive closure of schemas (mostly syntactic)

⇒nightmarish if 1, and 2. taken separately, horror marathon if considered together
Entities as Mediation

• Rising paradigm
  – Store information at the entity granularity
  – Integrate information by inter-linking entities

• Advantages?
  – Coarser granularity compared to keywords
    • More natural, e.g., brain functions similarly (or is it the other way around?)
  – Denormalized information compared to RDBMSs
    • Pre-computed joins
    • “Semantic” linking

• Drawbacks?
Example (1): Yahoo! Portal

**Parcel 104**
www.parcel104.com
(408) 970-8104
2700 Mission College Blvd, Santa Clara, CA

Make Reservations | Hours & More Info »

★★★★☆ (7)
7 Reviews - Parcel 104 is a real treat! The food is fantastic and it is nice not to be buried in tons of food that is too much for one meal. The... more

**Parcel 104 - Santa Clara, CA**
User Reviews | Photos | Write Review
- Yelp Rating: ★★★★★ 246
- Address: 2700 Mission College Blvd, Santa Clara, CA 95054, USA
- Phone: (408) 970-6104

www.yelp.com/biz/parcel-104-santa-clara - Cached

**Parcel 104, Santa Clara - Restaurant Reviews - TripAdvisor**
★★★★☆ - (408) 970-6104 - Santa Clara, CA 95054, 95054
#11 of 36 hotels in Sunnyvale, CA; Holiday Inn Express Mountain View (Town Center) ... provides unbiased reviews, articles, recommendations and opinions on Parcel 104 ...
www.tripadvisor.com/Restaurant_Review-g33046-d342112-... - Cached

**Sunnyvale American (New) Restaurants**
Parcel 104 is a diamond in the ruff for sure! This place is hidden at the Marriott ... Located in the heart of Downtown Sunnyvale, this place is really cute with super ...
www.yelp.com/c/sunnyvale-ca/newamerican - Cached
Example (2): Web Data Integration

http://data.semanticweb.org/person/philippe-cudre-mauroux
Focus on 4 Core Problems

1. Extracting entities from (HTML) text (*ZenCrowd*)

2. Searching for entities

3. Accessing entities (*DNS^3*)

4. Storing entities (*Diplodocus[RDF]*)
Extracting Entities

• Extracting entities from text is an old problem…
  – … and is extremely hard, esp. for machines

• Dozens of approaches have been suggested

• What if
  – We want to combine approaches / frameworks?
  – We want to leverage both human computations & algorithms?
ZenCrowd

• Exports entities from text using state-of-the-art techniques
• Uses sets of algorithmic matchers to match entities to online concepts
• Uses dynamic templating to create micro-matching-tasks and publish them on MTurk
• Combines both algorithmic and human matchers using probabilistic networks
“Black-Magic” Component

• Probabilistic network to integrate a priori & a posteriori information
  – Agreement of good turkers & algorithms
    • Learning process
  – Constraints
    • Unicity
    • Equality (SameAs)
  – Giant probabilistic graph
    • Instantiated selectively
Does it Work?

- Improves avg. prec. by 0.14 on average!
  - Minimal crowd involvement
  - Embarrassingly parallel problem
Searching for Entities

• How can end-users *reach* entities?
  ⇒ Keyword search
    • On their names or attributes
  – Obviously not ideal
    • BM25 on TREC 2011 AOR: $\text{MAP}=0.15$, $\text{P@10}=0.20$
    • Query extension, query completion or pseudo-relevance feedback yield comparable (or worse) results
Hybrid Entity Search

- Main idea: combine unstructured and structured search
  - Inverted index to locate first candidates
  - Graph queries to refine the results
    - Graph traversals (queries on object properties)
    - Graph neighborhoods (queries on data type properties)
Architecture

User

Entity Search Keyword Query

Graph-Enriched Results

3rd party search engines

Pseudo-Relevance Feedback

Query Annotation and Expansion

Ranking Functions

intermediate top-k results

WordNet

Structured Inverted Index

RDF Store

Graph Traversals (queries on object properties)

Neighborhoods (queries on datatype properties)

Final Ranking Function

Pseudo-Relevance Feedback

index()

query()
Query Refinement

• Finding the **right graph queries** to refine results is a real challenge
  - Which object property to follow? Transitive closures?
  - Which data type property to take into account? Scope?
Does it Work?

- Up to 25% improvement over best IR (stat. sign.)
- Very modest impact on latency (17%)
Whom to Ask for Entities?

- Void + SPARQL end-point
- DOA
- DNS [DNS\(^3\)]
- Same\_as service / Okkam IDs
- P2P Mesh of entities [idMesh]
- Downscaled registries?
How to Store Entities?

• Fundamental impedance mismatch between graphs of entities and...
  – N-ary / decomposition storage model
  – Inverted Indices
  – Key-value paradigms
Entity Storage

- Materialize the joins!
- Dense-pack the values
- Provide new indices

- Co-locate
- Co-locate
- Co-locate
DipLODocus: System Architecture

- **Workload**
  - Queries & Inserts
  - Results

- **Query Processor**
  - GetLists/GetClusters

- **Query Optimizer**

- **Template Manager**
  - Update Cluster

- **Cluster Manager**

- **Hash-Table**
  - URI
  - key

- **Templates**

- **Buffered operations**

- **Disks**
Main Idea - data structures

- Declarative Template
  - Molecule
  - List of Literals
  - Hash Table
Template Matching

(\text{Student032, FirstName, "Joe"}) \xrightarrow{match} 

\text{insert}

\text{hash("Joe")} \rightarrow \text{TID5: (cluster032)}

\text{Hash-Table}

\text{hash("Joe")}

\text{Cluster032}

\text{Template List 5}

\text{StudentInstance}

\text{TID: 1} \quad \text{StudentID}

\text{TID: 2} \quad \text{Student Class}

\text{TID: 3} \quad \text{Course Instance}

\text{TID: 4} \quad \text{LastName}

\text{TID: 5} \quad \text{FirstName}

\text{TID: 6} \quad \text{Is_a}

\text{Schema Template and Template IDs (TIDs)}
Hash Table

lexicographic tree to encode URIs

template based indexing

extremely compact lists of homologous nodes
Molecule Clusters

- extremely compact sub-graphs
- precomputed joins
List of Literals

- extremely compact list of sorted values
Basic operations - queries - triple patterns

?x type Student.
?x takesCourse Course0.

?x type Student.
?x takesCourse Course0.
?x takesCourse Course1.

=> intersection of sorted lists

=> intersection of sorted lists
Basic operations - queries, aggregates and analytics

?x type Student.
?x age ?y
filter (?y < 21)
Basic operations - queries - molecule queries

?a name 'Student1'.
?c ?d ?e.
Results - LUBM - 100 Universities

35 x faster for OLTP
300 x faster for OLAP

<table>
<thead>
<tr>
<th></th>
<th>q1</th>
<th>q2</th>
<th>q6</th>
<th>q8</th>
<th>q12</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigOwlim</td>
<td>3011.6</td>
<td>46.7</td>
<td>12.5</td>
<td>2343.1</td>
<td>698.3</td>
</tr>
<tr>
<td>Jena</td>
<td>75.1</td>
<td>17.8</td>
<td>5.4</td>
<td>828.9</td>
<td>128.5</td>
</tr>
<tr>
<td>virtuoso</td>
<td>25.3</td>
<td>37.9</td>
<td>109.9</td>
<td>599.4</td>
<td>4921.7</td>
</tr>
<tr>
<td>rdf3x</td>
<td>35.3</td>
<td>121.9</td>
<td>2.7</td>
<td>3.7</td>
<td>59.4</td>
</tr>
<tr>
<td>dipLODocus[RDF]</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- BigOwlim
- Jena
- virtuoso
- rdf3x
- dipLODocus[RDF]
Current Directions

• ZenCrowd: Web-scale Integration
• Entity Search: recommendation based on graph mining
• Entity registry: DOA / Downscaling for ad-hoc
• Diplodocus
  – Open-Source
  – Cloud Storage
  – Visualization (sensor data)
References

- ZenCrowd [WWW 2012]
- idMesh [WWW 2009]
- Hybrid Entity Search [SIGIR 2012]
- DNS^3 [ISWC 2011 demo]
- Downscaling Entity Registries [Downscale 2012]
- Diplodocus[RDF] [ISWC 2011]
- Graph Data Management for New Application Domains [VLDB 2011]
Upcoming Events

ISWC 2012
The 11th International Semantic Web Conference
November 11–15
Boston, USA

ISWC 2012 is the premier international forum for the Semantic Web / Linked Data Community. Here, scientists, industry specialists, and practitioners meet to discuss the future of practical, scalable, user-friendly, and game changing data.

Sponsors

SIMPDA 2012
IFIP Working Groups 2.6 and 2.12/12.4
18-20 June, 2012 - Campione d’Italia, Italy

--- Collocated with IEEE DEST-CEE 2012: http://sesar.dti.unimi.it/DEST2012 ---