Converging Web-Data and Database Data: Big - and Small Data via Linked Data

Moderation:
Fritz Laux, Reutlingen University, Germany

Panelists:
Andreas Schmidt, Karlsruhe Univ. of Applied Sciences and Karlsruhe Institute of Technology (KIT), Germany
Iztok Savnik, University of Primorska, Slovenia
Kiyoshi Nitta, Yahoo Japan Research, Japan
Hermann Kaindl, Vienna Univ. of Technology, Austria

Characterization of Web- and DB-Data

Web-Data
- Semi-structured: ad-hoc structure, verbose self-describing format → parsing (inefficient storing and retrieval)
- Size: largest data collection → unable to fit in central DBMS
- Quality: inconsistent, redundant → unreliable quality
- (approximate) Search is by keywords

DB-Data
- Structured: well defined formal data model → efficient storing and retrieval
- Size: limited to a specific domain
- Quality: consistent, reliable
- (exact) Query by logical expression

What should converge?
Characterization of Big and Small Data

**Big Data**
- Data that cannot be handled by a single system
  - Considering storage and processing power
- Mostly generated by machines or sensors
  - structured data, not text data \(\rightarrow\) industry 4.0
- Origin is unclear
  - John R. Mashey (sgi): Big Data ... Presentation on the next technology wave in 1998
- Data rich, but information poor

**Small Data**
- It is not the complement of Big Data!
- Small data connects people with timely, meaningful insights, organized and packaged to be accessible, understandable, and actionable for everyday tasks
  ([URL](http://smalldatagroup.com/2013/10/18/defining-small-data/), 2013)
  - Highly condensed and usable information/knowledge

What should converge?

Criteria for convergence

What should converge?

**Necessary conditions from the user’s perspective**
- Semantics
- User friendliness, quality
- Quality (performance, reliability, correctness)

**Necessary conditions from the developer’s perspective**
- Common data model, knowledge presentation
- Structure agnostic query/search
- Efficient query and reliable transaction technology

We need the convergence for all of the above!
Semantics: Model convergence

Tension field between schema strictness and structural complexity

To converge Web- and DB Data we need semantics and ultimately distill valuable knowledge out of the data!
INTRODUCTION

Resource description framework (RDF) data are widely used in the Internet and their volume is growing steadily. The linked open data (LOD) project promotes the acceleration of the accumulation of RDF data to provide freely accessible on-line resources.

40 billion triples

play an important role for query process efficiency
CLASSIFICATION OF RDF STORAGE MANAGERS

RDF storage managers in the local cache approach can be classified in accordance with several aspects.

\[ \text{RSM} (S, M) \]

PROPERTIES OF RDF STORAGE MANAGERS

<table>
<thead>
<tr>
<th>Manager</th>
<th>( T_s )</th>
<th>( I_s )</th>
<th>( Q_s )</th>
<th>( S_s )</th>
<th>( J_s )</th>
<th>( C_s )</th>
<th>( D_s )</th>
<th>( F_s )</th>
<th>( D_m )</th>
<th>( Q_m )</th>
<th>( S_m )</th>
<th>( A_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3store</td>
<td>v</td>
<td>S</td>
<td>U</td>
<td>R</td>
<td>R</td>
<td>T</td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>4store</td>
<td>v</td>
<td>S</td>
<td>U</td>
<td>o</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td>h</td>
<td>p</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>Virtuoso</td>
<td>v</td>
<td>G</td>
<td>S</td>
<td>Ulo</td>
<td>R</td>
<td>TA</td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>RDF-3X</td>
<td>v</td>
<td>6</td>
<td>S</td>
<td>Ul</td>
<td>o</td>
<td>R</td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Hexastore</td>
<td>v</td>
<td>6</td>
<td>o</td>
<td>Ul</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Apache Jena</td>
<td>p</td>
<td>S</td>
<td>Ulo</td>
<td>R</td>
<td>r</td>
<td>R</td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>SW-Store</td>
<td>h</td>
<td></td>
<td>Uo</td>
<td>c</td>
<td>m</td>
<td>c</td>
<td></td>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>BitMat</td>
<td>v</td>
<td>m</td>
<td>S</td>
<td>Ul</td>
<td>p</td>
<td>c</td>
<td></td>
<td></td>
<td>h</td>
<td>p</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>AllegroGraph</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p</td>
<td>m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hadoop/HBase</td>
<td>h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p</td>
<td>m</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHALLENGES

More varied values with S attributes than with M attributes

- Researches so far have succeeded in achieving good performances by developing single process technologies.
- While practical semantic web applications tend to process large-scale data sets, solutions based on data distribution parallelism have become more popular.

Caching techniques have not been researched that much

- Only Apache Jena and SW-Store reported confirming the efficiency of caching techniques.
- Technologies for automatic investigation and classification of processing queries might become important to utilize caching technologies.

Many researches have been carried out for developing efficient join algorithms with index structures

- This area has a long history in the research of database management systems.
- While the accumulated RDF data set is rapidly growing and SPARQL queries are basically constructed from joins of triple patterns, join operations will be applied more strongly in semantic web applications.

Most RDF storage managers can accept SPARQL queries

- SPARQL-based RDF storage managers rarely cause semantic mismatch due to the existence of RDF algebras described in the W3C recommendation.
- While OPTIONAL operator was introduced to make the query language convenient enough, efficient processing of such queries will be one of the most crucial challenges.
Advances on Converging WEB Data and Database Data: Big Data and Small Data via Linked Data

Andreas Schmidt

(1) Department of Informatics and Business Information Systems University of Applied Sciences Karlsruhe Moltkestraße 30 76133 Karlsruhe Germany

(2) Institute for Applied Sciences Karlsruhe Institute of Technologie PO-box 3640 76021 Karlsruhe Germany
Terminology

• What is ...
  • Big Data ?
  • Small Data ?
  • Linked Data ?
Terminology

• What is ...
  • Big Data
    • ... a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. [Snijders, C., Matzat, U., & Reips, U.-D. (2012). ‘Big Data’: Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 1-5]
    • Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers" [Jacobs, A. (6 July 2009). "The Pathologies of Big Data". ACMQueue]
  • Small Data
  • Linked Data
Terminology

• What is ...
  • Big Data
  • Small Data
    • “Small data is the amount of data you can conveniently store and process on a single machine, and in particular, a high-end laptop or server” (Rufus Pollok, Open Knowledge Foundation)
    • Small data connects people with timely, meaningful insights (derived from big data and/or “local” sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks (Alan Bonde)]
  • Linked Data
Terminology

- What is ... 
  - Big Data 
  - Small Data 
  - Linked Data 
    - ... describes a method of publishing structured data so that it can be inter-linked and become more useful [http://en.wikipedia.org/wiki/Linked_data] 
    - It builds upon standard Web technologies such as HTTP, RDF and URIs [] 
    - Origin: Open (Government) Data
Terminology

• Converging

\[
\begin{align*}
f(x) &= 2x^3 - 7x^2 + 12 \\
g(x) &= \frac{12x^5-14x + 3}{4x^2 + 2x}
\end{align*}
\]

\[
\lim_{x \to \infty} \left( \frac{f(x)}{g(x)} \right) = \frac{2}{3}
\]
• „Converging WEB Data and Database Data“

\[
\lim_{t \to \text{future}} \left( \frac{\text{web data}_t}{\text{database data}_t} \right) = \epsilon
\]

or, a „little“ be more formal ...

\[
\lim_{t \to \text{future}} \left( \frac{\text{op(web data}_t)}{\text{op(database data}_t)} \right) = \epsilon
\]

how does \text{op}(...) looks like?

what is \epsilon?
• „Big Data and Small Data via Linked Data“

\[
\text{linkify(Big Data)} \Rightarrow \text{linked data}
\]

\[
\text{linkify(Small Data)} \Rightarrow \text{linked data}
\]

\[
\text{Database Data} \Rightarrow \text{Web Data}
\]
• http://en.wikipedia.org/wiki/Linked_data
Data and Knowledge: One Man’s Opinion

Hermann Kaindl
Vienna University of Technology, Austria
Value Chain

- Knowledge Management
Miscellaneous

- „Datentechnik“ (data technology)
- From WWW to Semantic Web
- From GPS-driven navigation tools to Google cars
- Data in the Cloud
- Data looked outdated for a while, when everything seemed to be knowledge, but now data seem to be ubiquitous!
Thank you for your attention!
Current state of graph databases

Iztok Savnik
University of Primorska & Jožef Stefan Institute

Panel:
Big Data and Small Data via Linked Data
DBKDA, 2014
Terminology

- Linked data
  - Linked Open Data
- Open data
- Graph databases
- Knowledge bases
- Knowledge graphs
Wordnet

- Princeton's large lexical database of English.
  - Cognitive synonyms: synsets ≡ concepts
    - 117,000 synsets
  - Synsets are linked by:
    - conceptual-semantic relationships, and
    - lexical relationships.
    - Include definitions of synsets.
  - Main relationships:
    - Synonymy, hyponymy (ISA), meronymy (part-whole), antonymy
Linked Open Data

- Datasets are represented in RDF
  - Wikipedia, Wikibooks, Geonames, MusicBrainz, WordNet, DBLP bibliography
- Number of triples: 33 Giga ($10^9$) (2011)
- Governments:
  - USA, UK, Japan, Austria, Belgium, France, Germany, ...
- Active community

http://en.wikipedia.org/wiki/Open_Data
http://www.w3.org/LOD
Freebase

- Free, knowledge graph:
  - people, places and things,
  - 2,478,168,612 facts, 43,459,442 topics
- Semantic search engines are here!
Freebase

- Based on graphs:
  - nodes, links, types, properties, namespaces

- Google use of Freebase
  - Knowledge graph
  - Words become concepts
  - Semantic questions
  - Semantic associations
  - Browsing knowledge
  - Knowledge engine

- Available in RDF
YAGO

- 10 Mega (10^6) concepts
  - Max Planc Institute, Informatik
  - Accuracy of 95%

- Includes:
  - Wikipedia, WordNet, GeoNames
  - Links Wordnet to Wikipedia taxonomy (350K concepts)
  - Anchored in time and space
Wikidata

- Free knowledge base with 14,550,852 items
- Collecting structured data
- Properties of
  - person, organization, works, events, etc.
Wikidata

• Free knowledge base with 14,550,852 items
Cyc - knowledge base

- **Knowledge base**
  - Doug Lenat
  - Conceptual networks (ontologies)
  - Higher ontology, basic theories, specific theories
  - Predefined semantic relationships

- **Common sense reasoner**
  - Based on predicate calculus
  - Rule-based reasoning
Cyc

- Upper Ontology: Abstract Concepts
- Core Theories: Space, Time, Causality, ...
- Domain-Specific Theories
- Facts: Instances
Some conclusions

- There exist a variety of different dictionaries, properties, concepts, ...
  - Common definitions are not frequent
- There exist a variety of formats and models for knowledge and data representation
  - RDF is common data/knowledge model
- Senses of words are not represented