From the Web of Documents to the Web of Data

Marcelo Arenas
Pontificia Universidad Católica de Chile

ICWBD 2014, Goa, India
The Web of documents
But things have changed . . .
But things have changed . . .
A new opportunity: more structured queries

Who is the most cited researcher in area X in country Y?

The information is on the Web, the process can be automatized:

▶ Semantics: Interpret terms "most cited", "area X", . . .

▶ Distribution: Gather the needed pieces of information

▶ Heterogeneity: Integrate heterogeneous pieces of information
A new opportunity: more structured queries

Who is the most cited researcher in area $X$ in country $Y$?
A new opportunity: more structured queries

Who is the most cited researcher in area $X$ in country $Y$?

The information is on the Web, the process can be automatized:

▶ _Semantics_: Interpret terms “most cited”, “area $X$”, …

▶ _Distribution_: Gather the needed pieces of information

▶ _Heterogeneity_: Integrate heterogeneous pieces of information
We encounter similar challenges all around the Web
We encounter similar challenges all around the Web
We encounter similar challenges all around the Web

How to query distributed and heterogeneous semantic data?
Data sources keep getting bigger and bigger

Some of the known techniques are falling short.

We need to develop foundations and algorithms to take full advantage of the semantics of data at Web scale.
The Semantic Web
Semantic Web

“The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

[Tim Berners-Lee et al. 2001.]

Specific goals:
- Build a description language with standard semantics
  - Make semantics machine-processable and understandable
- Incorporate logical infrastructure to reason about resources
- W3C proposals: Resource Description Framework (RDF) and SPARQL
RDF in a nutshell

RDF is the framework proposed by the W3C to represent information in the Web:

- **URI vocabulary**
  - A URI is an atomic piece of data, and it identifies an abstract resource

- **Syntax based on directed labeled graphs**
  - URIs are used as node labels and edge labels

- **Schema definition language (RDFS):** Define new vocabulary
  - Typing, inheritance of classes and properties, ...

- **Formal semantics**
An example of an RDF graph: DBLP

```
: <http://dblp.l3s.de/d2r/resource/authors/>
conf: <http://dblp.l3s.de/d2r/resource/conferences/>
inPods: <http://dblp.l3s.de/d2r/resource/publications/conf/pods/>
swrc: <http://swrc.ontoware.org/ontology#>
dc: <http://purl.org/dc/elements/1.1/>
dct: <http://purl.org/dc/terms/>
```

![RDF graph diagram]

- `conf:pods`
- `swrc:series` to `inPods:2001`
- `dct:PartOf` to `inPods:FaginLN01`
- `dc:title` to "Optimal Aggregation ..."
- `dc:creator` to `:Amnon_Lotem`
- `dc:creator` to `:Moni_Naor`
- `dc:creator` to `:Ronald_Fagin`
An example of a URI

http://dblp.l3s.de/d2r/resource/conferences/pods
URI can be used for any abstract resource

http://dblp.l3s.de/d2r/page/authors/Ronald_Fagin
Why is this an interesting problem? Why is it challenging?

- RDF graphs can be interconnected
  - URIs should be dereferenceable

- Semantics of RDF is open world
  - RDF graphs are inherently incomplete
  - The possibility of adding optional information if present is an important feature

- Vocabulary with predefined semantics

- ...
Querying RDF: SPARQL

- SPARQL is the W3C recommendation query language for RDF (January 2008).
  - SPARQL is a recursive acronym that stands for SPARQL Protocol and RDF Query Language.

- SPARQL is a graph-matching query language.

- A SPARQL query consists of three parts:
  - Pattern matching: optional, union, filtering, . . .
  - Solution modifiers: projection, distinct, order, limit, offset, . . .
  - Output part: construction of new triples, . . .
SPARQL in a nutshell
SPARQL in a nutshell

SELECT ?Author
SPARQL in a nutshell

SELECT ?Author
WHERE
{
}
}
SPARQL in a nutshell

SELECT ?Author 
WHERE 
{
}

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
}
```
SPARQL in a nutshell

SELECT ?Author
WHERE
{
}
A SPARQL query consists of a:
A SPARQL query consists of a:

**Body:** Pattern matching expression
A SPARQL query consists of a:

Body: Pattern matching expression

Head: Processing of the variables
What are the challenges in implementing SPARQL?

SPARQL has to take into account the distinctive features of RDF:

- Should be able to extract information from interconnected RDF graphs
- Should be consistent with the open-world semantics of RDF
  - Should offer the possibility of adding optional information if present
- Should be able to properly interpret RDF graphs with a vocabulary with predefined semantics
Extracting information from interconnected RDF graphs

: <http://dblp.l3s.de/d2r/resource/authors/>
dbpedia: <http://dbpedia.org/resource/>
  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
  owl: <http://www.w3.org/2002/07/owl#>
yago: <http://dbpedia.org/class/yago>
dbo: <http://dbpedia.org/ontology/>

inPods:FaginLN01

dc:creator

:Ronald_Fagin

owl:sameAs

dbo:birthdayPlace
dbpedia:Oklahoma

dbo:birthPlace
dbpedia:Ronald_Fagin

rdf:type

yago:DatabaseResearchers

rdfs:subClassOf

yago:ResearchWorker
Dereferenceable URIs are the glue

http://dbpedia.org/resource/Ronald_Fagin
Querying interconnected RDF graphs

Retrieve the authors that have published in PODS and were born in Oklahoma:

```sparql
SELECT ?Author
WHERE
{
  ?Person owl:sameAs ?Author .
  ?Person dbo:birthdayPlace dbpedia:Oklahoma .
}
```
Retrieve the authors that have published in PODS, and their Web pages if this information is available:

```
SELECT ?Author ?WebPage
WHERE
{
  OPTIONAL { ?Author foaf:homePage ?WebPage . }
}
```
Taking into account vocabularies with predefined semantics

Retrieve the scientists that were born in Oklahoma and that have published in PODS:

SELECT ?Author
WHERE
{
  ?Author rdf:type yago:Scientist .
  ?Author dbo:birthPlace dbpedia:Oklahoma .
}
Taking into account vocabularies with predefined semantics

Retrieve the **scientists** that were born in Oklahoma and that have published in PODS:

```
  dbpedia:Ronald_Fagin
    rdf:type
    yago:DatabaseResearchers
      rdfs:subClassOf
      yago:Scientist
        rdf:type
        rdfs:subClassOf
        yago:ResearchWorker
          dbo:birthPlace
          dbpedia:Oklahoma
```
The Center for Semantic Web Research
The Center for Semantic Web Research

(funded by the Millennium Scientific Initiative)
Researchers

**Director**

Marcelo Arenas (PUC)  
*semantic Web, database theory*

**Deputy director**

Pablo Barcelo (UCHile)  
*graph databases, database theory*
Researchers

Director

Marcelo Arenas (PUC)  semantic Web, database theory

Deputy director

Pablo Barcelo (UChile)  graph databases, database theory

Associate researchers

Jorge Perez (UChile)  semantic Web, interoperability
Juan Reutter (PUC)  graph databases, interoperability
Claudio Gutierrez (UChile)  semantic Web, graph databases
Critical mass of young researchers

Young researchers

Renzo Angles (UTalca)  
Carlos Buil-Aranda (PUC)  
Aidan Hogan (UChile)  
Barbara Poblete (UChile) & Yahoo!  
Cristián Riveros (PUC)

Semantic Web  
Semantic Web  
Linked data  
Social networks  
Interoperability, automata

Graduate students

6 PhD & 3 postdocs
Strong international connections

IBM Almaden & Watson
U. of Oxford
U. of Texas at Austin
Rice U.
Microsoft Research
U. of Edinburgh
Polytechnic U. of Madrid
TU Vienna
U. of Bolzano
Digital Research Enterprise Institute (DERI)
Yahoo! Research
Our Proposal
Who is the most cited researcher in area X in country Y?
Who is the most cited researcher in area X in country Y?
Who is the most cited researcher in area X in country Y?

SELECT researcher
FROM DataWeb ...

∀x.∃y. S(x) → D(y)
Who is the most cited researcher in area X in country Y?

SELECT researcher FROM DataWeb ...

\( \forall x. \exists y. S(x) \rightarrow D(y) \)
Who is the most cited researcher in area X in country Y?

SELECT researcher FROM DataWeb ...

∀x.∃y.S(x) → D(y)
Who is the most cited researcher in area X in country Y?

SELECT researcher
FROM DataWeb ...

∀x.∃y.S(x) → D(y)
Who is the most cited researcher in area $X$ in country $Y$?

SELECT researcher FROM DataWeb ...

\[ \forall x. \exists y. S(x) \rightarrow D(y) \]
Identifying the right language for querying semantic data at Web scale

Who is the most cited researcher in area X in country Y?

```
SELECT researcher FROM DataWeb ...
∀x.∃y.S(x) → D(y)
```

Line 1

Researcher
A
B
...
Identifying the right language for querying semantic data at Web scale

- logic - finite model theory
- automata theory
- computational complexity

SELECT researcher FROM DataWeb ...

\( \forall x. \exists y. S(x) \to D(y) \)

Line 1
Who is the most cited researcher in area X in country Y?

SELECT researcher FROM DataWeb ...

∀x.∃y.\text{S}(x) \rightarrow D(y)

Line 1
Obtaining relevant information, efficiently

Who is the most cited researcher in area X in country Y?

SELECT researcher FROM DataWeb ...

∀x.∃y.S(x) → D(y)

Line 1

Line 2

Researcher

A

B

...
Obtaining relevant information, efficiently

- data structures, indexing
- query optimization
- (hyper)tree decomposition
- computational complexity
Obtaining relevant information, efficiently

Who is the most cited researcher in area \( X \) in country \( Y \)?

```sql
SELECT researcher
FROM DataWeb ...
∀x.∃y.\( S(x) \rightarrow D(y) \)
```

Line 1

Line 2

Researcher
A
B
...

A
Taking advantage of the structure of the data

Who is the most cited researcher in area $X$ in country $Y$?

SELECT researcher FROM DataWeb ...

$\forall x. \exists y. S(x) \rightarrow D(y)$
Taking advantage of the structure of the data

- graph theory
- network theory
- data dependency theory
Taking advantage of the structure of the data

Who is the most cited researcher in area X in country Y?

SELECT researcher FROM DataWeb ...

∀x.∃y. \( S(x) \rightarrow D(y) \)

Line 1

Line 2

Line 3

A

Researcher

A
B
...

...
Who is the most cited researcher in area X in country Y?

\[ \forall x. \exists y. S(x) \rightarrow D(y) \]

Line 1

SELECT researcher FROM DataWeb ...

Line 3

Line 2

Researcher

A
B
...

Line 4
Approximating answers when exact evaluation is infeasible

- graph theory
- approximation algorithms
- computational complexity
Who is the most cited researcher in area X in country Y?

\[ \forall x. \exists y. S(x) \rightarrow D(y) \]

Line 1

SELECT researcher FROM DataWeb ...

Line 2

Researcher
A
B
...

Line 4
(Some of) Our Projects
Publication of RDF Data

Translation of relational data into RDF

- Definition of a direct mapping, W3C standard: http://www.w3.org/TR/rdb-direct-mapping
- Study of fundamental notions such as information preservation, query preservation, ... [SAM12]
Publication of RDF Data

Translation of relational data into RDF

- Definition of a direct mapping, W3C standard: http://www.w3.org/TR/rdb-direct-mapping

- Study of fundamental notions such as information preservation, query preservation, ... [SAM12]

Generation of new RDF datasets from existing databases.

- Definition of a *declarative language* for HTML to RDF translation
Publication of RDF Data

Publication of public data

▶ Materialization of transparency law
  Design and (first) implementation of
  http://www.gobiernotransparentechile.cl and
  http://datos.gob.cl

▶ Scientific data from CONICYT: http://datoscientificos.cl
Study of the structure of RDF data

Study of the structuredness of RDF data [ADFKS14]
  ▶ Definition of a framework for specifying structuredness functions
  ▶ Study of the structure refinement problem

Study of the use of anonymous objects (blank nodes) in RDF data [HAMP]
  ▶ Reduction of the complexity of several reasoning problems
Storage of RDF data

Compression of RDF data [FMGPA13]

- HDT: defines header information, a dictionary, and the actual triples structure (http://www.rdfhdt.org)
- W3C submission: http://www.w3.org/Submission/2011/03
Study of Web query languages

Development of new benchmarks (http://www.ldbc.eu)
  ▶ To compare systems, and promote the development of new technologies

Study of the expressiveness of different query languages [AGP14, AP11, B13, BRV14, BLR14]
  ▶ What can and cannot be expressed in these languages
  ▶ What needs to be added to meet user requirements
  ▶ Study of new functionalities
Study of Web query languages

Development of query recommendation algorithms
  ▶ Definition of query extension and restriction
  ▶ Study of query logs (DBPedia, KEGG, ...)

Development of query evaluation algorithms

Indexing: Compression of RDF data [FMGPA13]

Incremental evaluation of SPARQL queries
  ► Development of algorithms, heuristics and data structures to efficiently updating answers to queries, in highly dynamic environments

Optimization and distribution of SPARQL queries [BHUV13,BACP13]
  ► Use of SPARQL endpoints
Development of query evaluation algorithms: MapReduce

- MapReduce has been a popular framework for parallel programming
- Very simple and useful language for engineers/programmers
- Good for optimizing massive parallel architectures
MapReduce drawbacks

- Not all problems are parallelizable

- What are the classes of problems that are optimizable in this framework?
Development of query evaluation algorithms: MapReduce

- Understand the computational power of the MapReduce framework
- Identify features of SPARQL that can be computed efficiently in this framework
- Extend/restrict SPARQL to exploit massive parallel architectures
Development of query approximation algorithms

Development meaningful notions of approximation [BLR13]

- Yield to efficient query evaluation algorithms
- Useful in applications in which data is massive and finding interconnection patterns is important (e.g. social networks, crime-detection networks, etc)
Thank you!
Bibliography


<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
<th>Title</th>
<th>Details</th>
</tr>
</thead>
</table>

