Optimization of Semantic Web Queries

Reinhard Pichler
Vienna University of Technology
Puebla, 19 May, 2013

*joint work with Andrés Letelier, Jorge Pérez, and Sebastian Skritek

Outline

1. Motivation
2. Tools for the Semantic Web
   2.1 RDF
   2.2 SPARQL
3. Role Model: Relational Algebra
4. Query plans and transformation rules for SPARQL
   4.1 Transformation rules for SPARQL
   4.2 Query Plans for SPARQL
5. Static Analysis of SPARQL Queries
6. Conclusion

Acknowledgement

I am very grateful to
- Marcelo Arenas and Jorge Pérez for providing me their slides from tutorials at WWW and PODS, respectively.
- Sebastian Skritek for his great help in creating this slide set.
Optimization of Semantic Web Queries

1. Motivation

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

Optimization of Semantic Web Queries

1. Motivation

2. Tools for the Semantic Web

RDF in a nutshell

W3C proposal framework for representing information on the web
- Extensible vocabulary based on URIs
  - URI = Uniform Resource Identifier
  - is an atomic piece of data; identifies an abstract resource.
- Abstract syntax based on directed labeled graph.
  - URIs are used as node labels and edge labels
- Formal semantics
- Existential variables as data values
- Support use of XML schema data types
- Schema definition language (RDFS): define new vocabulary
  - Typing, inheritance of classes and properties, ...
- Built-in vocabulary with fixed semantics (RDFS)
- Graph model where nodes may also be edge labels

Optimization of Semantic Web Queries

2. Tools for the Semantic Web

RDF in a nutshell

W3C proposal framework for representing information on the web
- Extensible vocabulary based on URIs
  - URI = Uniform Resource Identifier
  - is an atomic piece of data; identifies an abstract resource.
- Abstract syntax based on directed labeled graph.
  - URIs are used as node labels and edge labels
- Formal semantics
- Existential variables as data values
- Support use of XML schema data types
- Schema definition language (RDFS): define new vocabulary
  - Typing, inheritance of classes and properties, ...
- Built-in vocabulary with fixed semantics (RDFS)
- Graph model where nodes may also be edge labels

Outline

1. Motivation
2. Tools for the Semantic Web
   - 2.1 RDF
   - 2.2 SPARQL
3. Role Model: Relational Algebra
4. Query plans and transformation rules for SPARQL
   - 4.1 Transformation rules for SPARQL
   - 4.2 Query Plans for SPARQL
5. Static Analysis of SPARQL Queries
6. Conclusion
RDF formal model

\[(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)\] is called an RDF triple.

A set of RDF triples is called an RDF graph.
RDF: Example

```
{fcw name 'FC Wacker Innsbruck'. fcw league 'Bundesliga'.
f cw home 'tivoli'. tivoli poi 'stadium'. tivoli location innsbruck}
```

Reinhard Pichler
Puebla, 19 May, 2013

---

Querying RDF with triple patterns

- Triple patterns: RDF triple + variables

```
(?X, name, ?Name)
```

- The base case for RDF queries is a set of triple patterns

```
\{t_1, t_2, \ldots, t_k\}
```

This is called basic graph pattern (BGP).
Querying RDF with triple patterns

- Triple patterns: RDF triple + variables

\[(?X, \text{name}, ?Name)\]

- The base case for RDF queries is a set of triple patterns

\{t_1, t_2, \ldots, t_k\}.

This is called basic graph pattern (BGP).

Example

\{ (?X, \text{name}, ?Name), (?X, \text{home}, ?Stadium) \}

Mappings: building block for the semantics

Definition

A mapping is a partial function:

\[\mu : V \rightarrow (U \cup L)\]

Given a mapping \(\mu\) and a basic graph pattern \(P\):

- \(\text{dom}(\mu)\): the domain of \(\mu\).
- \(\mu(P)\): the set of triples obtained from \(P\) by replacing the variables according to \(\mu\).

The evaluation of a graph pattern results in a set of mappings.
The semantics of triple patterns

Given an RDF graph $G$ and a triple pattern $t$.

**Definition**

The evaluation of $t$ over $G$ is the set of mappings $\mu$ such that:
- $\mu$ has as domain the variables in $t$: $\text{dom}(\mu) = \text{var}(t)$
- $\mu$ makes $t$ to match the graph: $\mu(t) \in G$

**Example**

- triple pattern: $(\text{?X}, \text{location}, \text{?Y})$
- graph:
  - (tivoli, location, innsbruck)
  - (tivoli, poi, stadium)
  - (hanappi, location, wien)
- evaluation
  - $\mu_1$: tivoli innsbruck
  - $\mu_2$: hanappi wien
Compatible mappings

Definition

Mappings $\mu_1$ and $\mu_2$ are compatible if they agree in their common variables:

If $?X \in \text{dom}(\mu_1) \cap \text{dom}(\mu_2)$, then $\mu_1(?X) = \mu_2(?X)$.

Example

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>innsbruck</td>
<td>stadium</td>
<td>$M_1$</td>
</tr>
</tbody>
</table>

$\mu_1 :$

$\mu_2 :$

$\mu_3 :$

$\mu_1 \cup \mu_2 :$

$\mu_1 \cup \mu_3 :$

$\mu_2$ and $\mu_3$ are not compatible

Reinhard Pichler

Puebla, 19 May, 2013

Optimization of Semantic Web Queries

2. Tools for the Semantic Web

Compatible mappings

Definition

Mappings $\mu_1$ and $\mu_2$ are compatible if they agree in their common variables:

If $?X \in \text{dom}(\mu_1) \cap \text{dom}(\mu_2)$, then $\mu_1(?X) = \mu_2(?X)$.

Example

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>innsbruck</td>
<td>stadium</td>
<td>$M_1$</td>
</tr>
</tbody>
</table>

$\mu_1 :$

$\mu_2 :$

$\mu_3 :$

$\mu_1 \cup \mu_2 :$

$\mu_1 \cup \mu_3 :$

$\mu_2$ and $\mu_3$ are not compatible

Reinhard Pichler

Puebla, 19 May, 2013
Optimization of Semantic Web Queries

Querying RDF with 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”

- 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, boolean queries, ...

SPARQL in a nutshell

```sparql
SELECT ?Stadium 
WHERE 
{
  ?Club ex:league 'Bundesliga' .
}
```

A SPARQL query consists of a:

- **Body**: Pattern matching expression
**SPARQL in a nutshell**

```
SELECT ?Stadium
WHERE
  ?Club ex:league 'Bundesliga'.
}
```

A SPARQL query consists of a:

- **Body**: Pattern matching expression
- **Head**: Processing of the variable assignments

**Interesting features:**
- **Grouping**
- **Optional parts**

---

**SPARQL in a nutshell – SPARQL body**

**Body**: Pattern matching expression

```
{ { P1 
  P2 } 

{ P3 
  P4 } 

}
```

---

**SPARQL in a nutshell – SPARQL body**

**Body**: Pattern matching expression

```
{ { P1 
  P2 } 

OPTIONAL { P5 } } 

{ P3 
  P4 } 

OPTIONAL { P7 } 

}```
SPARQL in a nutshell – SPARQL body

Body: Pattern matching expression

Interesting features:
- Grouping
- Optional parts
- Nesting
- Union of patterns
- Filtering

{ { P1
  P2
  OPTIONAL { P5 } } 
{ P3
  P4
  OPTIONAL { P7
    OPTIONAL { P8 } } } } 

UNION
{ P9
  FILTER ( R ) }

---

SPARQL in a nutshell – SPARQL body

Body: Pattern matching expression

Interesting features:
- Grouping
- Optional parts
- Nesting
- Union of patterns
- Filtering
- Querying several sources

{ { P1
  P2
  OPTIONAL { P5 } } 
{ P3
  P4
  OPTIONAL { P7
    OPTIONAL { P8 } } } } 

UNION
{ P9
  FILTER ( R ) }
**SPARQL in a nutshell – SPARQL body**

**Body:** Pattern matching expression

Interesting features:
- Grouping
- Optional parts
- Nesting
- Union of patterns
- Filtering
- Querying several sources

We focus on the pattern matching expressions.

```sparql
{ { P1 
  P2 
  OPTIONAL { P5 } } 
{ P3 
  P4 
  OPTIONAL { P7 
    OPTIONAL { P8 } } } 
} UNION 
{ P9 
  FILTER ( R ) }
```

**A standard algebraic syntax**

- **Triple patterns:** triples including variables from a set \( V \)
  - \(?X ex:name 'wacker'\)
  - \((?X, name, 'wacker')\)

- **Graph patterns:** fully parenthesized algebra
  - \(\{ P1 P2 \}\)
  - \((P1 \text{ AND } P2)\)
  - \(\{ P1 \text{ OPTIONAL } P2 \}\)
  - \((P1 \text{ OPT } P2)\)
  - \(\{ P1 \text{ UNION } P2 \}\)
  - \((P1 \text{ UNION } P2)\)
  - \(\{ P1 \text{ FILTER } ( R ) \}\)
  - \((P1 \text{ FILTER } R)\)

**Setting considered here**

- **Restrictions**
  - Consider only the body of SPARQL queries, i.e., SPARQL graph patterns.
  - Consider only the operators AND and OPT.
  - Further restrictions will be introduced below.
Semantics of SPARQL: An example

Example

(tivoli, location, innsbruck)
(tivoli, poi, stadium)
(hanappi, location, wien)

( (?X, location, ?Y) OPT (?X, poi, ?E) )
Semantics of SPARQL: An example

Example

(tivoli, location, innsbruck)
(tivoli, poi, stadium)
(hanappi, location, wien)

\[( (?X, \text{location}, ?Y) \text{OPT} ( ?X, \text{poi}, ?E) ) \]

<table>
<thead>
<tr>
<th>?X</th>
<th>?Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>innsbruck</td>
</tr>
<tr>
<td>hanappi</td>
<td>wien</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>?X</th>
<th>?E</th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>stadium</td>
</tr>
</tbody>
</table>
SPARQL in a nutshell

Definition
The EVALUATION problem for SPARQL graph patterns:
INPUT: An RDF dataset $G$, a graph pattern $P$ and a mapping $\mu$.
QUESTION: Is $\mu \in \llbracket P \rrbracket_G$? ($\llbracket P \rrbracket_G$ = the solutions of $P$ over $G$)

Observation
- High complexity due to unrestricted occurrences of OPT-operator.
- Arbitrary use of OPT-operator is highly unintuitive.

Example

graph:
(fcw, home, tivoli), (fcw, league, 'Bundesliga'),
(hanappi, location, wien)

graph pattern:

result:
{{?C → fcw, ?S → tivoli}}
Optimization of Semantic Web Queries  
2. Tools for the Semantic Web

Observation

- High complexity due to unrestricted occurrences of OPT-operator.
- Arbitrary use of OPT-operator is highly unintuitive.

Example

graph:
(fcw, home, tivoli), (fcw, league, ‘Bundesliga’),
(hanappi, location, wien)

graph pattern:

result:
{{ ?C → fcw, ?S → tivoli }}

Reinhard Pichler
Puebla, 19 May, 2013

Well-designed graph patterns

Definition

A query in the AND-FILTER-OPT fragment of SPARQL is well-designed if for every OPT in the pattern:

( ··········· ( P OPT Q ) ··········· )

if a variable occurs inside Q
Well-designed graph patterns

| Definition | A query in the AND-FILTER-OPT fragment of SPARQL is well-designed if for every OPT in the pattern:
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( ··········· ( P OPT Q ) ··········· ) ↑ ↑ ↑ ↑</td>
</tr>
<tr>
<td></td>
<td>if a variable occurs inside Q and anywhere outside the OPT operator, then the variable must also occur inside P.</td>
</tr>
</tbody>
</table>

Example

(\(C\), home, \(S\)) OPT (\(C\), league, \(L\)) OPT (\(S\), location, \(P\))
Well-designed graph patterns

**Definition**

A query in the AND-FILTER-OPT fragment of SPARQL is well-designed if for every OPT in the pattern:

\[
( \ldots ( P \text{ OPT } Q ) \ldots )
\]

if a variable occurs inside \( Q \) and anywhere outside the \( \text{OPT} \) operator, then the variable must also occur inside \( P \).

**Example**

\[
\]

Reinhard Pichler

Puebla, 19 May, 2013

How common are well-designed patterns?

What are real SPARQL queries like?

[Picalausa and Vansummeren, 2011]

- DBpedia query log: 623,000 queries without UNION operator

Reinhard Pichler

Puebla, 19 May, 2013
How common are well-designed patterns?

What are real SPARQL queries like? [Picalausa and Vansummeren, 2011]
- DBpedia query log: 623,000 queries without UNION operator
- 52% of these queries are well-designed

Non-well-designed queries from DBpedia usually look unnatural.

They often seem to be bugs rather than intended features.

Well-designed graph patterns

Theorem (Pérez, Arenas, Gutierrez 2009)

\textbf{EVALUATION} is coNP-complete for well-designed graph pattern expressions constructed by using only \textbf{AND} and \textbf{OPT} operators.

The story so far

- Syntax of RDF
  - RDF triples \((s, p, o)\)
  - underlying data model: labelled, directed graph

- Basic querying of RDF
  - RDF triple patterns
  - Mappings

- SPARQL
  - SPARQL graph patterns
  - restriction to operators OPT and AND
  - well-designed SPARQL
  - complexity of SPARQL
What comes next

- **Ultimate goal**: optimization of SPARQL queries
- learn from successful methods from the relational world
- first steps to apply analogous methods to SPARQL

### Outline

1. **Motivation**
2. **Tools for the Semantic Web**
   2.1 RDF
   2.2 SPARQL
3. **Role Model: Relational Algebra**
4. Query plans and transformation rules for SPARQL
   4.1 Transformation rules for SPARQL
   4.2 Query Plans for SPARQL
5. Static Analysis of SPARQL Queries
6. Conclusion

### Relational Algebra

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_\alpha$</td>
<td>Selection</td>
</tr>
<tr>
<td>$\pi_{\vec{A}}$</td>
<td>Projection</td>
</tr>
<tr>
<td>$\rho_{A \leftarrow B}$</td>
<td>Renaming</td>
</tr>
<tr>
<td>$R_1 \times R_2$</td>
<td>Cross Product</td>
</tr>
<tr>
<td>$R_1 \bowtie R_1$</td>
<td>Join</td>
</tr>
<tr>
<td>$R_1 \setminus R_2$</td>
<td>Set Difference</td>
</tr>
<tr>
<td>$R_1 \cup R_2$</td>
<td>Union</td>
</tr>
<tr>
<td>$R_1 \cap R_2$</td>
<td>Intersection</td>
</tr>
</tbody>
</table>

**Example**

**Relations:** location(stadium, city), home(club, stadium), league(club, league).

**SQL Query:**

```
SELECT city
FROM location, home, league
WHERE location.stadium = home.stadium AND
home.club = league.club AND league.league = 'Bundesliga';
```

**Relational Algebra:**

```
\pi_{\text{location.city}}(\sigma_{P}(\text{location} \times \text{home} \times \text{league}))
```

with

```
P := \text{location.stadium} = \text{home.stadium} \land \text{home.club} = \text{league.club}
\land \text{league.league} = 'Bundesliga'
```
Relational Algebra

Example

Query Plan:

\[ \pi_{\text{location.city}} \]

\[ \sigma_{\text{location.stadium}=\text{home.stadium} \land \text{home.club}=\text{league.club} \land \text{league.league}='\text{Bundesliga}' } \]

\[ \times \]

location

home

league

\[ \times \]

Reinhard Pichler

Puebla, 19 May, 2013
Relational Algebra

Example

Consider the following database instance:

<table>
<thead>
<tr>
<th>location:</th>
<th>home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>fcw</td>
</tr>
<tr>
<td>innsbruck</td>
<td>tivoli</td>
</tr>
<tr>
<td>hanappi</td>
<td>scr</td>
</tr>
<tr>
<td>wien</td>
<td>hanappi</td>
</tr>
<tr>
<td>allianz</td>
<td>fcb</td>
</tr>
<tr>
<td>munchen</td>
<td>allianz</td>
</tr>
<tr>
<td>camp nou</td>
<td>barca</td>
</tr>
<tr>
<td>barcelona</td>
<td>barcelona</td>
</tr>
<tr>
<td>old trafford</td>
<td>manu</td>
</tr>
<tr>
<td>manchester</td>
<td>old trafford</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>league:</th>
</tr>
</thead>
<tbody>
<tr>
<td>fcw ‘Bundesliga’</td>
</tr>
<tr>
<td>scr ‘Bundesliga’</td>
</tr>
<tr>
<td>fcb ‘Fb-Bundesliga’</td>
</tr>
<tr>
<td>barca ‘PD’</td>
</tr>
<tr>
<td>manu ‘PL’</td>
</tr>
</tbody>
</table>

Recall the Query Plan:
Optimization of Semantic Web Queries

3. Role Model: Relational Algebra

Recall the Query Plan:

\[ \pi_{\text{location.city}} \]
\[ \sigma_{\text{location.stadium}=\text{home.stadium} \land \text{home.club}=\text{league.club} \land \text{league.league}='Bundesliga'} \]

\[ \times \]

\[ \times \]

location: 5  
home: 5  
league: 5  

\[ \times \]

location: 5  
home: 5  
league: 5

Reinhard Pichler  
Puebla, 19 May, 2013
Relational Algebra

Example
Recall the Query Plan:

\[ \pi_{\text{location.city}} \]
\[ \sigma_{\text{league.league}='Bundesliga'} \]
\[ \bowtie_{\text{location.stadium}='home.stadium'} 5 \]
\[ \text{league: 5} \]

location: 5  home: 5

Example
Query Plan:

\[ \pi_{\text{location.city}} \]
\[ \sigma_{\text{league.league}='Bundesliga'} \]
\[ \bowtie_{\text{location.stadium}='home.stadium'} 5 \]
\[ \text{league: 5} \]

location: 5  home: 5
$\pi_{\text{location.city}}$

$\bowtie R_1 \text{club} = \text{league.club}$

$\sigma_{\text{league.league} = \text{'Bundesliga'} }$

- fcw
- scr

$\sigma_{\text{league.league} = \text{'Bundesliga'} }$

- fcw
- scr

$\sigma_{\text{league.league} = \text{'Bundesliga'} }$

- fcw
- scr

Lessons learned from the relational world

- an appropriate data structure for representing queries: simply use expression trees
- a semantics to directly use these trees for query evaluation: consider the trees as Query Execution Plans
- equivalence preserving transformations
Lessons learned from the relational world

We need...

- an appropriate data structure for representing queries: simply use expression trees
- a semantics to directly use these trees for query evaluation: consider the trees as Query Execution Plans
- equivalence preserving transformations

Conclusion: We also need an appropriate algebra for SPARQL!

Normal form of well-designed graph patterns

OPT normal form [Pérez, Arenas, Gutierrez 2009]

- a basic graph pattern is in OPT-normal form
- $(P_1 \text{ OPT } P_2)$ is in OPT normal form if $P_1$ and $P_2$ are.

Intuition: no OPT-operator is in the scope of an AND-operator.
Normal form of well-designed graph patterns

**OPT normal form [Pérez, Arenas, Gutierrez 2009]**

- a basic graph pattern is in OPT-normal form
- \((P_1 \text{ OPT } P_2)\) is in OPT normal form if \(P_1\) and \(P_2\) are.

Intuition: no OPT-operator is in the scope of an AND-operator.

**Rewrite rules to obtain OPT normal form**

- \((P_1 \text{ AND } (P_2 \text{ OPT } P_3))\) \(\rightarrow\) \((P_1 \text{ AND } P_2) \text{ OPT } P_3\)
- \(((P_1 \text{ OPT } P_2) \text{ AND } P_3)\) \(\rightarrow\) \((P_1 \text{ AND } P_3) \text{ OPT } P_2\)

**Property [Pérez, Arenas, Gutierrez 2009]**

For well-designed SPARQL patterns:
\(((P_1 \text{ OPT } P_2) \text{ OPT } P_3)\) \(\equiv\) \(((P_1 \text{ OPT } P_3) \text{ OPT } P_2)\)

**Nesting of OPTs**

\[((P_1 \text{ OPT } P_2) \text{ OPT } P_3)\]
- First try to extend solutions of \(P_1\) to solutions of \(P_2\).
- Try to extend the resulting mappings to solutions of \(P_3\).
- As seen above, we could change this order.
- \(\Rightarrow P_2\) and \(P_3\) are on the “same level”.

\((P_1 \text{ OPT } (P_2 \text{ OPT } P_3))\)
- First try to extend solutions of \(P_2\) to solutions of \(P_3\).
- Then try to extend the solutions of \(P_1\) to these mappings.
- Solutions of \(P_1\) are extended to solutions of \(P_3\) only if we first can extend the solutions of \(P_1\) to \(P_2\).
- \(\Rightarrow P_2\) is “one level above” \(P_3\).

**Pattern trees**

Consider a pattern in OPT-normal form

\(((?S, \text{poi}, \text{stadium}) \text{ \text{AND} } (?C, \text{home}, ?S) \text{ \text{AND} } (?C, \text{league, ‘Bundesliga’})\))

\(\text{OPT}\)

\(((?S, \text{location, ?L}) \text{ \text{AND} } (?U, \text{category, university}) \text{ \text{AND} } (?U, \text{location, ?L})\))

\(\text{OPT}\)

\(((?A, \text{location, ?L}) \text{ \text{AND} } (?S, \text{location, ?L}) \text{ \text{AND} } (?A, \text{poi, ?I})\))

\(\text{OPT}\)

\(((?S, \text{offers, visitor_center})\))

\(\text{OPT}\)

\(((?S, \text{ticketprice}, ?P))\)

\(\text{OPT}\)

\(((?S, \text{phone, ?N}))\)

---

Reinhard Pichler
Puebla, 19 May, 2013
Pattern trees

Consider a pattern in OPT-normal form

\[
\begin{align*}
&((?S, \text{poi, stadium}) \land (?C, \text{home, ?S}) \land (?C, \text{league, 'Bundesliga'})\bigg) \\
&\text{OPT} \\
&\bigg((?S, \text{location, ?L}) \land (?U, \text{category, university}) \land (?U, \text{location, ?L})\bigg) \\
&\text{OPT} \\
&\bigg((?A, \text{location, ?L}) \land (?S, \text{location, ?L}) \land (?A, \text{poi, ?I})\bigg) \\
&\text{OPT} \\
&\bigg((?S, \text{offers, visitor_center}) \land (?S, \text{ticketprice, ?P}) \land (?S, \text{phone, ?N})\bigg) \\
\end{align*}
\]

Pattern trees as query plans

\[
\begin{align*}
&((?S, \text{poi, stadium}), (?C, \text{home, ?S}), (?C, \text{league, 'Bundesliga'})\bigg) \\
&\text{OPT} \\
&\bigg((?S, \text{location, ?L}), (?U, \text{category, university}), (?U, \text{location, ?L})\bigg) \\
&\text{OPT} \\
&\bigg((?A, \text{location, ?L}), (?S, \text{location, ?L}), (?A, \text{poi, ?I})\bigg) \\
&\text{OPT} \\
&\bigg((?S, \text{ticketprice, ?P}), (?S, \text{phone, ?N})\bigg) \\
\end{align*}
\]

Pattern trees as query plans – Bottom Up Evaluation

Example RDF graph

(louvre, location, paris), (louvre, poi, museum), (eiffelt., location, paris), (eiffelt., poi, steel),
(g.cathedral, location, galway), (g.cathedral, poi, church), (santaluciahill, location, santiago),
(santaluciahill, poi, hill), (albertina, location, wien), (albertina, poi, museum),
(st.jakob, location, innsbruck), (st.jakob, poi, church), (UPEC, category, university),
(UPEC, location, paris), (NUIG, category, university), (NUIG, location, galway),
(PUC, category, university), (PUC, location, santiago), (TUW, category, university),
(TUW, location, wien), (UIBK, category, university), (UIBK, location, innsbruck),
(tivoli, poi, stadium), (tivoli, location, innsbruck), (fcw, home, tivoli), (fcw, league, 'Bundesliga'),
(hanappi, poi, stadium), (hannapi, location, wien), (scr, home, hanappi),
(scr, league, 'Bundesliga'), (pappel, poi, stadium), (pappel, location, mattersburg),
(svm, home, pappel), (svm, league, 'Bundesliga'), (louvre, offers, visitor_center),
(hanappi, offers, visitor_center), (eiffelt., offers, visitor_center), (louvre, ticketprice, '39'),
(albertina, ticketprice, '11'), (hannapi, ticketprice, '6'), (tivoli, ticketprice, '20'),
(PUC, home, nr. 1), (TUW, home, nr. 2), (albertina, phone, nr. 3), (louvre, phone, nr. 4),
(hannapi, phone, nr. 5), (pappel, phone, nr. 6),

Pattern trees as query plans – Bottom Up Evaluation

\[
\{ (?S, poi, stadium), (?C, home, ?S), (?C, league, ‘Bundesliga’) \}
\]

\[
\{ (?S, location, ?L), (?U, category, university), (?U, location, ?L) \}
\]

\[
\{ (?A, location, ?L), (?S, location, ?L), (?A, poi, ?I) \}
\]

\[
\{ (?S, offers, visitor_center) \}
\]

\[
\{ (?S, ticketprice, ?P) \}
\]

\[
\{ (?S, phone, ?N) \}
\]

\[
\{ (?S, phone, ?N) \}
\]

\[
\{ (?S, phone, ?N) \}
\]

<table>
<thead>
<tr>
<th>?S</th>
<th>?N</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUC</td>
<td>nr. 1</td>
</tr>
<tr>
<td>TUW</td>
<td>nr. 2</td>
</tr>
<tr>
<td>albertina</td>
<td>nr. 3</td>
</tr>
<tr>
<td>louvre</td>
<td>nr. 4</td>
</tr>
<tr>
<td>hannapi</td>
<td>nr. 5</td>
</tr>
<tr>
<td>pappel</td>
<td>nr. 6</td>
</tr>
</tbody>
</table>
Pattern trees as query plans – Bottom Up Evaluation

\[ (?S, ticketprice, ?P) \]

\[ (?S, poi, stadium), (?C, home, ?S), (?C, league, ‘Bundesliga’) \]

\[ (?S, location, ?L), (?U, category, university), (?U, location, ?L) \]

\[ (?A, location, ?L), (?S, location, ?L), \]

\[ (?S, offers, visitor_center) \]

\[ (?S, ticketprice, ?P) \]

\[ (?S, phone, ?N) \]

\[ \]
Pattern trees as query plans – Bottom Up Evaluation

\[
\begin{array}{c}
(S, offers, visitor\_center) \\
(S, offers, visitor\_center) \\
(S, offers, visitor\_center)
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
louvre & '39'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
albertina & '11'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
hannapi & '6'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
tivoli & '20'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
louvre & '39'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
ahannapi & '6'
\end{array}
\]

\[
\begin{array}{cc}
?S & ?P \\
eiffelt. &
\end{array}
\]
Pattern trees as query plans – Bottom Up Evaluation

\{
(\text{?}S, \text{poi}, \text{stadium}), (\text{?}C, \text{home}, \text{?}S), (\text{?}C, \text{league}, \text{’Bundesliga’})
\}

\{
(\text{?}S, \text{location}, \text{?}L), (\text{?}U, \text{category}, \text{university}), (\text{?}U, \text{location}, \text{?}L)
\}

\{
(\text{?}A, \text{location}, \text{?}L), (\text{?}S, \text{location}, \text{?}L), (\text{?}A, \text{poi}, \text{?}I)
\}

\{
(\text{?}S, \text{offers}, \text{visitor\_center})
\}

\{
(\text{?}S, \text{ticketprice}, \text{?}P)
\}

\{
(\text{?}S, \text{phone}, \text{?}N)
\}

Reinhard Pichler
Puebla, 19 May, 2013

Optimization of Semantic Web Queries
4. Query plans and transformation rules for SPARQL

Pattern trees as query plans – Bottom Up Evaluation

\{
(\text{?}S, \text{location}, \text{?}L), (\text{?}U, \text{category}, \text{university}), (\text{?}U, \text{location}, \text{?}L)
\}

\{
(\text{?}A, \text{location}, \text{?}L), (\text{?}S, \text{location}, \text{?}L), (\text{?}A, \text{poi}, \text{?}I)
\}

\{
(\text{?}S, \text{ticketprice}, \text{?}P)
\}

\{
(\text{?}S, \text{phone}, \text{?}N)
\}

Reinhard Pichler
Puebla, 19 May, 2013
Pattern trees as query plans – Bottom Up Evaluation

\[
\begin{array}{ccc}
\end{array}
\]

\[
\{ (?A, location, ?L),
(?S, location, ?L),
(?A, poi, ?I) \}
\]

Pattern trees as query plans – Bottom Up Evaluation

\[
\{ (?S, poi, stadium),
(?C, home, ?S),
(?C, league, ‘Bundesliga’) \}
\]

\[
\{ (?S, location, ?L),
(?U, category, university),
(?U, location, ?L) \}
\]

\[
\{ (?A, location, ?L),
(?S, location, ?L),
(?A, poi, ?I) \}
\]

\[
\{ (?S, offers, visitor_center) \}
\]

\[
\{ (?S, ticketprice, ?P) \}
\]

\[
\{ (?S, phone, ?N) \}
\]

\[
\{ (?A, location, ?L),
(?S, location, ?L),
(?A, poi, ?I) \}
\]

\[
\{ (?S, location, ?L),
(?U, category, university),
(?U, location, ?L) \}
\]

\[
\{ (?A, location, ?L),
(?S, location, ?L),
(?A, poi, ?I) \}
\]

\[
\{ (?A, location, ?L),
(?S, location, ?L),
(?A, poi, ?I) \}
\]

\[
\{ (?S, location, ?L),
(?U, category, university),
(?U, location, ?L) \}
\]
Pattern trees as query plans – Bottom Up Evaluation

(louvre, location, paris), . . . (g.cathedral, location, galway), . . .
(UPEC, category, university), (UPEC, location, paris), (NUIG, category, university),
(NUIG, location, galway), . . . (UIBK, category, university), (UIBK, location, innsbruck), .
(tivoli, location, innsbruck), ..........................................................

?(S, location, ?L),
(?U, category, university),
(?U, location, ?L)
Pattern trees as query plans – Bottom Up Evaluation


\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)\}

\{(?S, offers, visitor_center)\}

\{(?S, ticketprice, ?P)\} \{(?S, phone, ?N)\}


\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)\}

\{(?S, offers, visitor_center)\}

\{(?S, ticketprice, ?P)\} \{(?S, phone, ?N)\}
Pattern trees as query plans – Bottom Up Evaluation

\[
\text{\{(S, poi, stadium), (C, home, S), (C, league, 'Bundesliga')\}}
\]

\[
\text{\{(S, location, ?L), (U, category, university), (U, location, ?L)\}}
\]

\[
\text{\{(A, location, ?L), (S, location, ?L), (S, offers, visitor_center)\}}
\]

\[
\text{\{(S, ticketprice, ?P)\}}
\]

\[
\text{\{(S, phone, ?N)\}}
\]
Pattern trees as query plans – Bottom Up Evaluation

Result of Bottom-up evaluation

Example

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tivoli</td>
<td>fcw</td>
<td>innsbruck</td>
<td>UIBK</td>
<td>tivoli</td>
<td>stadium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hanappi</td>
<td>scr</td>
<td>wien</td>
<td>TUW</td>
<td>hanappi</td>
<td>stadium</td>
<td>'6'</td>
<td>nr. 5</td>
</tr>
<tr>
<td>pappel</td>
<td>svm</td>
<td>mattersburg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reinhard Pichler
Puebla, 19 May, 2013

Alternative Evaluation: Top-Down Evaluation

Observation
- Bottom-up evaluation starts at the leaves.
- Mappings computed for some node can possibly not be extended upwards.
- The natural direction for extending graph patterns is downwards.
- Idea: choose a top-down evaluation instead.

Pattern trees as query plans – Top Down Evaluation


{{(?S, location, ?L), (?U, category, university), (?U, location, ?L)}}

{{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)}}
{{(?S, offers, visitor_center)}}

{{(?S, ticketprice, ?P)}}
{{(?S, phone, ?N)}}
Pattern trees as query plans – Top Down Evaluation

\{(?S, poi, stadium), (?C, home, ?S), (?C, league, 'Bundesliga')\}

\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)\}

\{(?S, offers, visitor_center)\}

\{(?S, ticketprice, ?P)\} \{(?S, phone, ?N)\}

Pattern trees as query plans – Top Down Evaluation

\{(?S, poi, stadium), (?C, home, ?S), (?C, league, 'Bundesliga')\}

\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)\}

\{(?S, offers, visitor_center)\}

\{(?S, ticketprice, ?P)\} \{(?S, phone, ?N)\}
Pattern trees as query plans – Top Down Evaluation

\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\begin{align*}
\text{?S} & \quad \text{?C} \\
\text{tivoli} & \quad \text{fcw} \\
\text{hanappi} & \quad \text{scr} \\
\text{pappel} & \quad \text{svm} \\
\text{?T} & \quad \text{?I} & \quad \text{?S} & \quad \text{?L} & \quad \text{?U} & \quad \text{?C} \\
\text{tivoli} & \quad \text{innsbruck} & \quad \text{UIBK} & \quad \text{fcw} \\
\text{hanappi} & \quad \text{wien} & \quad \text{TUW} & \quad \text{scr} \\
\text{pappel} & \quad \text{svm} \\
\end{align*}
Pattern trees as query plans – Top Down Evaluation


\{(?S, location, ?L), (?U, category, university), (?U, location, ?L)\}

\{(?A, location, ?L), (?S, location, ?L), (?A, poi, ?I)\}

\{(?S, ticketprice, ?P)\} \{(?S, phone, ?N)\}

{(?S, offers, visitor_center)}

\begin{tabular}{|c|c|c|c|c|c|}
\hline
\hline
Tivoli & Stadium & Tivoli & Innsbruck & UIBK & Fcw \\
St. Jakob & Church & Tivoli & Innsbruck & UIBK & Fcw \\
Hanappi & Stadium & Hanappi & Wien & TUW & Scr \\
Albertina & Museum & Hanappi & Wien & TUW & Scr \\
pappel & & & & & svm \\
\hline
\end{tabular}
Pattern trees as query plans – Top Down Evaluation

\[
\{(S, ticketprice, P)\}.
\]

\begin{array}{|c|c|c|c|c|c|}
\hline
A & I & S & L & U & C \\
\hline
\hline
tivoli & stadium & tivoli & innsbruck & UIBK & fcw \\
\hline
st.jakob & church & tivoli & innsbruck & UIBK & fcw \\
\hline
hanappi & stadium & hanappi & wien & TUW & scr \\
\hline
albertina & museum & hanappi & wien & TUW & scr \\
\hline
\end{array}

\[
\{(S, ticketprice, P)\}.
\]

\begin{array}{|c|c|c|c|c|c|}
\hline
A & I & S & L & U & C \\
\hline
\hline
tivoli & stadium & tivoli & innsbruck & UIBK & fcw \\
\hline
st.jakob & church & tivoli & innsbruck & UIBK & fcw \\
\hline
hanappi & stadium & hanappi & wien & TUW & scr \\
\hline
albertina & museum & hanappi & wien & TUW & scr \\
\hline
\end{array}

\[
\text{(hanappi, ticketprice, '6')},
\]

\[
\{(S, phone, N)\}.
\]

\begin{array}{|c|c|c|c|c|c|}
\hline
A & I & S & L & U & C \\
\hline
\hline
tivoli & stadium & tivoli & innsbruck & UIBK & fcw \\
\hline
st.jakob & church & tivoli & innsbruck & UIBK & fcw \\
\hline
hanappi & stadium & hanappi & wien & TUW & scr \\
\hline
albertina & museum & hanappi & wien & TUW & scr \\
\hline
\end{array}

\[
\text{(hanappi, phone, nr. 5)},
\]

\[
\{(S, phone, N)\}.
\]
Pattern trees as query plans – Top Down Evaluation

The story so far

- RDF
- SPARQL
- restriction to well-designed SPARQL
- OPT normal form
- Pattern trees
  - representation of well-designed SPARQL queries
  - useful as query execution plans
  - "top-down" instead of "bottom-up" evaluation

Next step: transformations of pattern trees
Transformation Rules for Pattern Trees

\[
\{ (\text{?S, poi, stadium}), (\text{?C, home, ?S}), (\text{?C, league, ‘Bundesliga’}) \}
\]

\[
\{ (\text{?S, location, ?L}), (\text{?U, category, university}), (\text{?U, location, ?L}) \}
\]

\[
\{ (\text{?A, location, ?L}), (\text{?S, location, ?L}), (\text{?A, poi, ?I}) \}
\]

\[
\{ (\text{?S, offers, visitor_center}) \}
\]

\[
\{ (\text{?S, ticketprice, ?P}) \} \quad \{ (\text{?S, phone, ?N}) \}
\]

Reinhard Pichler  
Puebla, 19 May, 2013
Transformation Rules for Pattern Trees


\textit{R1} If a triple pattern \( t \) belongs to node \( n \) and to a descendant \( n' \) of \( n \), then delete \( t \) from \( n' \).

\textit{R2} If node \( n \) does not introduce any new variable w.r.t. its ancestors, then push copies of \( n \) into its children.

\textit{R3} If there is a homomorphism from node \( n \) to the branch from the root to the parent of \( n \), then merge node \( n \) with its parent.
Outline

1. Motivation
2. Tools for the Semantic Web
   2.1 RDF
   2.2 SPARQL
3. Role Model: Relational Algebra
4. Query plans and transformation rules for SPARQL
   4.1 Transformation rules for SPARQL
   4.2 Query Plans for SPARQL
5. Static Analysis of SPARQL Queries
6. Conclusion

Optimization of Semantic Web Queries

Query Containment and Equivalence

Definition

- $Q_1 \subseteq Q_2$ iff $Q_1(D) \subseteq Q_2(D)$ for all databases $D$.
- $Q_1 \equiv Q_2$ iff $Q_1(D) = Q_2(D)$ for all databases $D$.

Equivalence:

- **Optimization**: Replacing a query by an equivalent, but “better” one.

Containment:

- $Q_1 \equiv Q_2$ iff $Q_1 \subseteq Q_2$ and $Q_2 \subseteq Q_1$
- Relationship to query answering, e.g. BCQs ask if the database is contained in the query

SPARQL: Subsumption

Definition

Let $P_1$ and $P_2$ be two SPARQL graph patterns. We say that $P_1$ is subsumed by $P_2$ ($P_1 \subseteq P_2$) if, for every RDF graph $G$, every solution of $P_1$ over $G$ can be extended to a solution of $P_2$ over $G$.

In case of SPARQL, subsumption is the more natural measure.

Example

- $P_1 = (?C, n, ?N)$
- $P_2 = (?C, n, ?N) \text{OPT} (?C, h, ?H)$
- $[P_1]_G = \{\mu\}$ with $\mu = \{?C \rightarrow fcw, ?N \rightarrow 'wacker'\}$
- $[P_2]_G = \{\mu'\}$ with $\mu' = \{?C \rightarrow fcw, ?N \rightarrow 'wacker', ?H \rightarrow tivoli\}$

In case of SPARQL, subsumption is the more natural measure.
Relationship: Subsumption and Equivalence

**Theorem**

Let $P_1$ and $P_2$ be two SPARQL graph patterns. Then $P_1 \equiv P_2$ if and only if $P_1 \sqsubseteq P_2$ and $P_2 \sqsubseteq P_1$.

**Proof idea.**

⇒) Trivial

⇐) Assume $\mu \in [P_1]_G$ with $\mu \notin [P_2]_G$.

- By $P_1 \sqsubseteq P_2$, there is an extension $\mu_2$ of $\mu$, s.t. $\mu_2 \in [P_2]_G$.
- Since $\mu \notin [P_2]_G$, $\mu_2$ is a proper extension of $\mu$.
- By $P_2 \sqsubseteq P_1$, there is an extension $\mu_1$ of $\mu_2$, s.t. $\mu_1 \in [P_1]_G$.
- Hence, $\mu_1$ is a proper extension of $\mu$, s.t. $\mu_1 \in [P_1]_G$.
- Contradiction to $\mu \in [P_1]_G$.

Complexity of Subsumption

**Definition**

Let SUBSUMPTION be the following problem:

INPUT: Two SPARQL graph patterns $P_1$ and $P_2$.

QUESTION: Does $P_1 \sqsubseteq P_2$ hold?

**Theorem**

For well-designed SPARQL graph patterns, SUBSUMPTION is $\Pi_2^p$-complete.

Complexity of Equivalence

**Definition**

Let EQUIVALENCE be the following problem:

INPUT: SPARQL graph patterns $P_1$ and $P_2$.

QUESTION: Does $P_1 \equiv P_2$ hold?

**Theorem**

For well-designed SPARQL graph patterns, EQUIVALENCE is NP-complete.
Complexity of Equivalence

Proof idea.

- pattern trees in normal form w.r.t. rules R1, R2, and R3.
- necessary conditions for equivalence
  - the two pattern trees must have the same set of variables
  - they must have identical roots
- define an appropriate notion of homomorphism
- give an equivalence criterion based on the existence of certain homomorphisms (between nodes with common new variables)

Conclusion

Evaluation of SPARQL graph patterns

- pattern trees: an abstract representation formalism for well-designed SPARQL graph patterns
- top-down evaluation of pattern trees
- optimization and transformation rules on pattern trees

Fundamental computational problems

- SUBSUMPTION is \( \Pi_2^P \)-complete for well-designed SPARQL graph patterns.
- EQUIVALENCE is \( \text{NP} \)-complete for well-designed SPARQL graph patterns.

Future Work

Extend work on well-designed SPARQL

- Identify further transformation rules
- Verify usefulness of transformation rules for optimization

Consider extensions of well-designed SPARQL

- Extend representation formalism and transformation rules to more operators, especially UNION and FILTER
- Consider SPARQL 1.1 features, e.g., entailment regimes
- Consider bag semantics