In this color!
In this color!
I want to do a project and a thesis.
I like Semantic Web.
What can be a good research topic? Finding a good research topic (and good supervisor) is also part of research!
Werner: “Hi, we are doing research about completeness reasoning on databases”
Me: “Why not also on the Semantic Web?”
Motivation


Completeness statement about the IMDB data source

Quentin Tarantino was the character Mr. Brown

Fariz Darari (EMCL Workshop 2014)  Completeness Reasoning @ Linked Data
Motivation

I want to find out all movies directed by and starring Tarantino.

Syd

LMDdb is complete for Tarantino movies and their actors.

DBpedia is complete for Tarantino movies.
Hmmm..for realizing my goal, what should I learn?

Aha, I need to learn Semantic Web, I think the Semantic Web Technologies course offered by FUB would be useful for me.

Aha, I also need to learn existing work on completeness reasoning for databases, I guess this paper\(^1\) is worth to read!

---

\(^1\)Completeness of Queries over Incomplete Databases by Simon and Werner
Story: Google (and your supervisors) are your Googles, ask them!

About 15,000,000 results (0.27 seconds)

**Database completeness** - IBM
pic.dhe.ibm.com/infocenter/.../c_sysadm_db_completeness.html - IBM
Database completeness. The standard backup and restore by using the nzbackup and nzrestore commands provide transactionally consistent, automated ...

**[PDF]** **Completeness of Queries over Incomplete Databases** - VLDB ...
We develop techniques to conclude the completeness of query answers from information about the completeness of parts of a generally incomplete database.

**What is completeness** constraint - Wiki Answers
wiki.answers.com › ... › Computer Programming › Database Programming
The Completeness Constraint addresses the issue of whether or not an ... concurrency control and why you think its important in database enviroment.

**[PDF]** **Integrity = Validity + Completeness** - Department of Comp cs.gmu.edu/~ami/research/.../tods89.pdf - George Mason University - Cited by 161 - Related articles
Completeness reasoning on databases has been investigated.

Hmm.. but databases are different than Linked Data, what should be adjusted?

Well, in Linked Data, instead of databases, we have RDF data sources.

Well, in Linked Data, instead of SQL, we have SPARQL as a query language.

Well, Linked Data also is more open, more heterogeneous and federated.

Well, Linked Data also has ontologies.

Then, I have to discuss these issues with my supervisors!
Incomplete Data Source

Quentin Tarantino is missing..

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:runtime</td>
<td>5940 (xsd:double)</td>
</tr>
</tbody>
</table>
An incomplete data source of Tarantino movies, $G_{dbp} = (G^a_{dbp}, G^i_{dbp})$: 

- Movie
- reservoir Dogs
- killBill
- pulp Fiction
- desperado
- tarantino

- type
- director
- actor
An incomplete data source of Tarantino movies, $G_{dbp} = (G^a_{dbp}, G^i_{dbp})$: 

![Diagram](image)

- **Movie**
  - **reservoir Dogs**
  - **killBill**
  - **pulp Fiction**
  - **desperado**

**Nodes:** 
- *reservoir Dogs* 
- *killBill* 
- *pulp Fiction* 
- *desperado* 
- *tarantino*

**Edges:**
- Type
- Director
- Actor

[Image: Incomplete Data Source](image)
An incomplete data source of Tarantino movies, $G_{dbp} = (G^a_{dbp}, G^i_{dbp})$:
An **incomplete data source** is a pair of two graphs, 

$$\mathcal{G} = (G^a, G^i), \text{ where } G^a \subseteq G^i.$$ 

We call $G^a$ the **available** graph and $G^i$ the **ideal** graph.
To express that a source is complete for all the triples about movies directed by Tarantino, we use the statement

\[ C_{\text{dir}} = \text{Compl}((?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino}) \mid \emptyset), \]
Completeness Statements: Examples

To express that a source is complete for all the triples about movies directed by Tarantino, we use the statement

$$C_{dir} = \text{Compl}((?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino}) \mid \emptyset)$$

To express that a source is complete for all triples about actors in movies directed by Tarantino, we use

$$C_{act} = \text{Compl}((?m, \text{actor}, ?a) \mid (?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino}))$$
Let $P_1$ be a non-empty BGP (Basic Graph Pattern) and $P_2$ a BGP.

A completeness statement is defined as

$$\text{Compl}(P_1 \mid P_2)$$

where we call $P_1$ the pattern and $P_2$ the condition of the statement.
Satisfaction of Completeness Statements

To the statement

\[ C = \text{Compl}(P_1 \mid P_2), \]

we associate the **CONSTRUCT** query

\[ Q_C = (P_1, P_1 \cup P_2). \]

Then, we say:

\( C \) is **satisfied** by an IDS \( \mathcal{G} = (G^a, G^i) \), written \( \mathcal{G} \models C \), if

\[ \llbracket Q_C \rrbracket_{G^i} \subseteq G^a. \]
Completeness Statements

\[ C_{dir} = \text{Compl}(\langle ?m, \text{type}, \text{Movie} \rangle, \langle ?m, \text{director}, \text{tarantino} \rangle \mid \emptyset) \]

Question: \( G_{dbp} \models C_{dir} ? \)
Completeness Statements

\[ C_{dir} = \text{Compl}( (?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino}) \mid \emptyset) \]

Question: \( G_{\text{dbp}} \models C_{dir} \)?

Yes, because \( [ Q_{C_{dir}} ] G^i_{\text{dbp}} \subseteq G^a_{\text{dbp}} \).

Dashed lines represent what are missing in DBpedia.
\[ C_{act} = Compl((?m, \text{actor}, ?a) | (?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino})) \]

**Question:** \( G_{dbp} \models C_{act} ? \)
Completeness Statements

\[ C_{act} = \text{Compl}((?m, \text{actor}, ?a) \mid (?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino})) \]

Question: \( G_{dbp} \models C_{act} \)?

No, because among the results of \( \left[ \text{Q}_{C_{act}} \right] G_{dbp}^i \), there is \( \text{(reservoirDogs, actor, tarantino)} \) not in \( G_{dbp}^a \).
Completeness Statements in RDF

\[ C_{act} = Compl( (?m, \text{actor}, ?a) \mid (?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino}) ) \]
$Q_{dir} = (\{?m\}, \{(?m, \text{type}, \text{Movie}), (?m, \text{director}, \text{tarantino})\})$
Query Completeness: Example

\[ Q_{\text{dir}} = (\{?m\}, \{(?m, \text{type, Movie}), (?m, \text{director, tarantino})\}) \]

Then, \( [Q_{\text{dir}}]_{G_{dbp}^i} = [Q_{\text{dir}}]_{G_{dbp}^a} \), and therefore \( G_{dbp} \models \text{Compl}(Q_{\text{dir}}) \).
Let $Q$ be a `SELECT` query. We write $\text{Compl}(Q)$ to say that $Q$ is complete. An incomplete data source $\mathcal{G} = (G^a, G^i)$ satisfies $\text{Compl}(Q)$, written $\mathcal{G} \models \text{Compl}(Q)$ if and only if $[Q]_{G^i} = [Q]_{G^a}$. 

**Definition**
Theorems are formal representations of goals you want to achieve (remember the first slides about motivation)

In fact..all the definitions and framework introduced before are actually to well-define these theorems

Hmm..I am sure these theorems should hold!

Then, I have to discuss these issues with my supervisors!
Let $C$ be a set of completeness statements and $Q$ a \texttt{SELECT} query. We say that $C$ entails the completeness of $Q$, written $C \models \text{Compl}(Q)$, if any incomplete data source satisfying $C$ also satisfies $\text{Compl}(Q)$. 
Transfer Operator

For any set $\mathcal{C}$ of completeness statements and a graph $G$, we define the **transfer operator** $T_{\mathcal{C}}$:

$$T_{\mathcal{C}}(G) = \bigcup_{C \in \mathcal{C}} [Q_C]_G$$

Prototypical Graph

Let $Q = (W, P)$ be a SELECT query. The prototypical graph $\tilde{P}$ is the graph resulting from the mapping of variables in $P$ to fresh, unique IRIs.
Transfer Operator

For any set $\mathcal{C}$ of completeness statements and a graph $G$, we define the **transfer operator** $T_\mathcal{C}$:

$$T_\mathcal{C}(G) = \bigcup_{\mathcal{C} \in \mathcal{C}} [Q_\mathcal{C}]_G$$

Prototypical Graph

Let $Q = (W, P)$ be a *SELECT* query. The **prototypical graph** $\tilde{P}$ is the graph resulting from the mapping of variables in $P$ to fresh, unique IRIs.
Completeness of Basic Queries

**Theorem**

Let $\mathcal{C}$ be a set of completeness statements and $Q = (W, P)$ a basic query. Then,

$$\mathcal{C} \models \text{Compl}(Q) \quad \text{if and only if} \quad \tilde{P} = T_C(\tilde{P}).$$
Story: Start Doing The Hardcore Part - Proving Theorems

- Argh, I got stuck
- This proving theorems stuff always haunts me!
- But:

 Because you are Excellent Marvelous Champion Legendary (aka EMCL)
Get the proof idea first

Generate some concrete examples of the problems with their solutions from the theorem you want to prove

Discuss with your supervisors

Now, start to prove in detail

There is a problem on this, and that, then refine your proof idea, go back to the first point

Start doing the above points until (hopefully :) you are able to prove it
Consider the set of completeness statements

$$C_{dir,act} = \{ C_{dir}, C_{act} \}$$

and the query

$$Q_{dir+act} = (\{ ?m \}, P_{dir+act})$$

where

$$P_{dir+act} = \{ (?m, type, Movie), (?m, director, tarantino), (?m, actor, tarantino) \}.$$
Completeness Reasoning: Example

Consider the set of completeness statements

\[ C_{dir,act} = \{ C_{dir}, C_{act} \} \]

and the query

\[ Q_{dir+act} = (\{ ?m \}, P_{dir+act}) \]

Then,

\[ \tilde{P}_{dir+act} = \{ (c?m, type, Movie), (c?m, director, tarantino), (c?m, actor, tarantino) \} \]
Completeness Reasoning: Example

\[ C_{\text{dir}, \text{act}} = \{ C_{\text{dir}}, C_{\text{act}} \} \]

\[ Q_{\text{dir}+\text{act}} = (\{ ?m \}, P_{\text{dir}+\text{act}}) \]

Therefore,

\[ T_{C_{\text{dir}, \text{act}}}(\tilde{P}_{\text{dir}+\text{act}}) \]
\( C_{dir,act} = \{ C_{dir}, C_{act} \} \)

\( Q_{dir+act} = (\{ ?m \}, P_{dir+act}) \)

Therefore,

\[
\mathcal{T}_{C_{dir,act}}(\tilde{P}_{dir+act}) = [Q_{C_{dir}}]\tilde{P}_{dir+act} \cup [Q_{C_{act}}]\tilde{P}_{dir+act}
\]
$C_{dir,act} = \{ C_{dir}, C_{act} \}$

$Q_{dir+act} = (\{ ?m \}, P_{dir+act})$

Therefore,

$$\mathcal{T}_{C_{dir,act}} (\tilde{P}_{dir+act}) = \left[ Q_{C_{dir}} \right] \tilde{P}_{dir+act} \cup \left[ Q_{C_{act}} \right] \tilde{P}_{dir+act} = \left\{ \left( c?m, \text{type, Movie} \right), \left( c?m, \text{director, tara} \right), \right\}$$
Completeness Reasoning: Example

\[ C_{dir, act} = \{ C_{dir}, C_{act} \} \]

\[ Q_{dir+act} = (\{ ?m \}, P_{dir+act}) \]

Therefore,

\[
\mathcal{T}_{C_{dir, act}}(\tilde{P}_{dir+act}) \\
= \left[ Q_{C_{dir}} \right] \tilde{P}_{dir+act} \cup \left[ Q_{C_{act}} \right] \tilde{P}_{dir+act} \\
= \{ (c?m, type, Movie), (c?m, director, tara), (c?m, actor, tara) \} 
\]
Completeness Reasoning: Example

\[ C_{dir, act} = \{ C_{dir}, C_{act} \} \]

\[ Q_{dir+act} = (\{ ?m \}, P_{dir+act}) \]

Therefore,

\[
\begin{align*}
\mathcal{T}_{C_{dir, act}}(\tilde{P}_{dir+act}) &= \left[ Q_{C_{dir}} \right] \tilde{P}_{dir+act} \cup \left[ Q_{C_{act}} \right] \tilde{P}_{dir+act} \\
&= \{ (c?m, type, Movie), (c?m, director, tara), (c?m, actor, tara) \} \\
&= \tilde{P}_{dir+act}.
\end{align*}
\]

Thus, \( C_{dir, act} \models \text{Compl}(Q_{dir+act}) \)
The framework can also be applied to:

- **DISTINCT** Queries: with set semantics
- **OPT** Queries: eg. get all people and in case they are not single, get also their spouse
- Queries with RDFS Data Sources: incorporating ontologies
Federated Framework

Can I get a complete answer?

```
SELECT ?m ?l
WHERE { ?m rdf:type s:Movie .
  ?m s:director dbp:tarantino .
  ?m fbo:likes ?l }
```
Can check the completeness of a subset of SPARQL queries depending on the completeness statements of a single data source.

Developed in Java using the Apache Jena library.

Takes three inputs:
- Completeness statements in RDF format
- A SPARQL query
- (optional) an RDFS ontology

Available at [http://rdfcorner.wordpress.com/](http://rdfcorner.wordpress.com/).
I have got all the results
But, I haven’t started writing the thesis yet
Nooo, I will miss the deadline :(
I have got all the results
And, I have started writing the thesis since the very beginning (right after I have studied the literature)
I think I can manage to meet the deadline :)
Aha, I think our research results can be also useful if shown to the Semantic Web community, why not to try submit it at ISWC 2013? – Another research story :)

Fariz Darari (EMCL Workshop 2014)
We developed a theoretical framework for the expression of completeness statements about data sources, from which we can ensure query completeness.

We provide a compact RDF representation of completeness statements, which can be embedded in VoID descriptions, and implemented the framework in CoRner.

We are interested in studying completeness reasoning with more expressive queries and OWL 2.
Terima kasih, grazie, danke!!