Ontology-Based Semantic Search on the Web

Thomas Lukasiewicz

University of Oxford, UK
thomas.lukasiewicz@cs.ox.ac.uk

Outline

Motivation

- Web search is a key technology of the Web.
- Web search is about to change radically with the development of the Semantic Web as a more powerful future Web:
  - ...an extension of the current Web by standards and technologies that help machines to understand the information on the Web, to support richer discovery, data integration, navigation, and automation of tasks.
- Very recent joint initiative of Google, Microsoft, and Yahoo to add meaning to Web pages to aid search.
- The development of a new semantic search technology for the Web, called semantic search on the Web, is currently a very hot topic, both in Web-related companies and in academic research:
  - There is a fastly growing number of commercial and academic semantic search engines for the Web.

Key Ideas of *Serene*

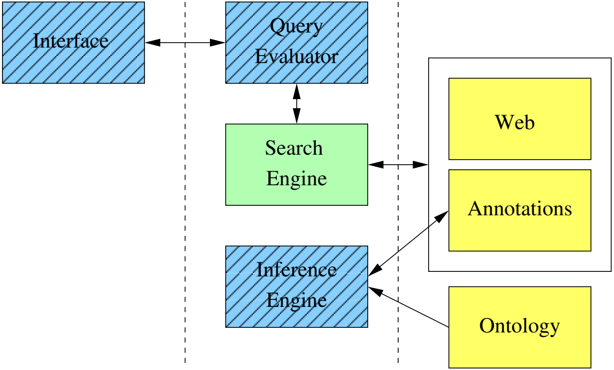
- Connect the information on existing Web pages with background ontological knowledge.
- Mapping Web pages/objects to a knowledge base relative to an ontology; vertical vs. general search.
- Make current search engines more "semantic" / "intelligent" (adds meaning and structure to Web pages and queries).
- Semantic search on the Web on top of standard Web search:
  - can immediately be applied to the existing Web (and not only to the future Semantic Web), and
  - it can be done with existing Web search technology (and so does not require completely new technologies).
- More complex search queries and more precise answers; reasoning over the contents of Web pages.

Examples

- When searching for a movie, one may be interested in movies that were produced by a US company before 1999 and had a French director.
- When buying a house in a town, one may be interested in large house selling companies within 50 miles of that town, existing for at least 15 years, and not known to be blacklisted by a consumer organization in the last 5 years.
- When searching for “laptop”, then one is looking for laptops or synonyms / related concepts (such as “notebook”), but also for special kinds of laptops that are not synonyms / related concepts, such as e.g. IBM/Lenovo ThinkPads.

- A search for “president of the USA” should also return Web pages that contain “George W. Bush” (who was one of the presidents of the USA according to some background ontology).
- A search for “the president of the USA on September 11, 2001” should return Web pages mentioning “George W. Bush” (who was the president of the USA on September 11, 2001, according to some background ontology).
- When searching for Web pages about the first president of the USA, “Washington”, semantic annotations and background knowledge allow us to restrict our search to Web pages that are actually about Washington as the name of the president, and so to ignore, e.g., Web pages about the state or town.

System Architecture



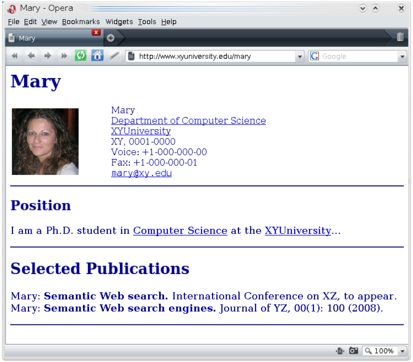
Semantic Annotations

We assume semantic annotations to standard Web pages and to objects on standard Web pages:

- user-defined: starting to be widely available for a large class of Web resources, especially with the Web 2.0;
- automatically learned from the Web pages and the objects to be annotated;
- automatically extracted from Web pages via user-defined rules (i.e., mapping Web pages/objects to an ontological knowledge base).

Example

A Web page i_1 may contain information about a Ph.D. student i_2, called Mary, and two of her papers, namely, a conference paper i_3 entitled "*Semantic Web search*" and a journal paper i_4 entitled "*Semantic Web search engines*" and published in 2008.



Annotation for the Web page encodes that it mentions Mary and the two papers:

$$\mathcal{A}_{i_1} = \{\mathit{contains}(i_1, i_2), \mathit{contains}(i_1, i_3), \mathit{contains}(i_1, i_4)\}.$$

Annotation for Mary may encode that she is a Ph.D. student with the name Mary and the author of the papers i_3 and i_4:

$$\mathcal{A}_{i_2} = \{\mathit{PhDStudent}(i_2),\ \mathit{name}(i_2, \text{``}\mathit{mary}\text{''}), \mathit{isAuthorOf}(i_2, i_3), \mathit{isAuthorOf}(i_2, i_4)\}.$$

Annotation for the paper i_3 may encode that i_3 is a conference paper and has the title "*Semantic Web search*":

$$\mathcal{A}_{i_3} = \{\mathit{ConferencePaper}(i_3), \mathit{title}(i_3, \text{``}\mathit{Semantic\ Web\ search}\text{''})\}.$$

Annotation for the paper i_4 may encode that i_4 is a journal paper, authored by Mary, has the title "*Semantic Web search engines*", was published in 2008, and has the keyword "RDF":

$$\mathcal{A}_{i_4} = \{\mathit{JournalPaper}(i_4),\ \mathit{hasAuthor}(i_4, i_2),\ \mathit{title}(i_4, \text{``}\mathit{Semantic\ Web\ search\ engines}\text{''}), \mathit{yearOfPublication}(i_4, 2008), \mathit{keyword}(i_4, \text{``RDF''})\}.$$

Inference Engine

Using a background ontology, these semantic annotations are then further enhanced in an offline inference step, where the Inference Engine adds all properties that can be deduced / induced from the semantic annotations and the ontology.

The resulting (completed) semantic annotations are then published as Web pages, so that they can be searched by standard Web search engines.

Example

An ontology may contain the knowledge that all journal and conference papers are also articles, that conference papers are not journal papers, and that "is author of" is the inverse relation to "has author", which is formally expressed by the axioms

$$\mathit{ConferencePaper} \sqsubseteq \mathit{Article},\ \mathit{JournalPaper} \sqsubseteq \mathit{Article},$$
$$\mathit{ConferencePaper} \sqsubseteq \neg \mathit{JournalPaper},$$
$$\mathit{isAuthorOf}^{-} \sqsubseteq \mathit{hasAuthor},\ \mathit{hasAuthor}^{-} \sqsubseteq \mathit{isAuthorOf}.$$

Using this ontological background knowledge, we can derive from the above annotations that the two papers i_3 and i_4 are also articles, and are both authored by Mary.

These searchable completed semantic annotations of (objects on) standard Web pages are published as HTML Web pages with pointers to the respective object pages.

www.xyuniversity.edu/mary/an1.html

```
<html>
<body>
www.xyuniversity.edu/mary <br>
WebPage i1 <br>
contains i2 <br>
contains i3 <br>
contains i4 <br>
</body>
</html>
```

www.xyuniversity.edu/mary/an2.html

```
<html>
<body>
www.xyuniversity.edu/mary <br>
PhDStudent i2 <br>
name mary <br>
isAuthorOf i3 <br>
isAuthorOf i4 <br>
</body>
</html>
```

www.xyuniversity.edu/mary/an3.html

```
<html>
<body>
www.xyuniversity.edu/mary <br>
Article i3 <br>
ConferencePaper i3 <br>
hasAuthor i2 <br>
title Semantic Web search <br>
</body>
</html>
```

www.xyuniversity.edu/mary/an4.html

```
<html>
<body>
www.xyuniversity.edu/mary <br>
Article i4 <br>
JournalPaper i4 <br>
hasAuthor i2 <br>
title Semantic Web search engines <br>
yearOfPublication 2008 <br>
keyword RDF <br>
</body>
</html>
```

Query Evaluator

The Query Evaluator reduces each Semantic Web search query in an online step to a sequence of standard Web search queries on standard Web and annotation pages, which are then processed by a standard Web Search Engine, assuming standard Web and annotation pages are appropriately indexed.

The Query Evaluator also collects the results and re-transforms them into a single answer which is returned to the user.

Example

Semantic Web search query, one may ask for all Ph.D. students who have published an article in 2008 with RDF as a keyword, which is formally expressed as follows:

$$Q(x) = \exists y\,(\mathit{PhDStudent}(x) \wedge \mathit{isAuthorOf}(x, y) \wedge \mathit{Article}(y) \wedge \mathit{yearOfPublication}(y, 2008) \wedge \mathit{keyword}(y, \text{“}\mathit{RDF}\text{”}))\,.$$

This query is transformed into the two queries $Q_1 = \mathit{PhDStudent}$ AND $\mathit{isAuthorOf}$ and $Q_2 = \mathit{Article}$ AND *“yearOfPublication* 2008” AND *“keyword* RDF”, which are both submitted to a standard Web search engine.

The result of the original query Q is then constructed from the results of the two queries Q_1 and Q_2.

Ontology Languages

As underlying ontology language, we use the tractable description logic *DL-Lite*$_{\mathcal{A}}$, which adds datatypes to a restricted combination of the tractable description logics *DL-Lite*$_{\mathcal{F}}$ and *DL-Lite*$_{\mathcal{R}}$. All these description logics belong to the *DL-Lite* family.

The *DL-Lite* description logics are a class of restricted description logics for which the main reasoning tasks are possible in polynomial time in general and some of them even in LOGSPACE in the data complexity. The *DL-Lite* description logics are fragments of OWL and the most common tractable ontology languages in the Semantic Web.

Example

Sets of atomic concepts, atomic roles, atomic attributes, individuals, and data values:

$$\begin{aligned}
\mathbf{A} &= \{\mathit{Scientist}, \mathit{Article}, \mathit{ConferencePaper}, \mathit{JournalPaper}\},\\
\mathbf{R}_A &= \{\mathit{hasAuthor}, \mathit{isAuthorOf}\},\\
\mathbf{R}_D &= \{\mathit{name}, \mathit{title}, \mathit{yearOfPublication}\},\\
\mathbf{I} &= \{i_1, i_2\},\\
\mathbf{V} &= \{\text{“}\mathit{mary}\text{”}, \text{“}\mathit{Semantic\ Web\ search}\text{”}, 2008\}.
\end{aligned}$$

TBox $\mathcal{T}$ contains the subsequent axioms, which informally express that (i) conference and journal papers are articles, (ii) conference papers are not journal papers,...:

$$\begin{aligned}
&\mathit{ConferencePaper} \sqsubseteq \mathit{Article},\ \mathit{JournalPaper} \sqsubseteq \mathit{Article},\\
&\mathit{ConferencePaper} \sqsubseteq \neg\mathit{JournalPaper},\\
&\exists \mathit{isAuthorOf} \sqsubseteq \mathit{Scientist},\ \exists \mathit{isAuthorOf}^- \sqsubseteq \mathit{Article},\\
&\mathit{isAuthorOf}^- \sqsubseteq \mathit{hasAuthor},\ \mathit{hasAuthor}^- \sqsubseteq \mathit{isAuthorOf},\\
&(\text{funct } \mathit{hasFirstAuthor}).
\end{aligned}$$

ABox $\mathcal{A}$ contains the following axioms, which express that the individual i_1 is a scientist whose name is "*mary*" and who is the author of article i_2, which is entitled "*Semantic Web search*" and has been published in the year 2008:

$$Scientist(i_1),\ name(i_1, \text{"}mary\text{"}), isAuthorOf(i_1, i_2),$$
$$Article(i_2),\ title(i_2, \text{"}Semantic\ Web\ search\text{"}),$$
$$yearOfPublication(i_2, 2008).$$

Querying for all scientists who published an article in 2008 can be expressed by the following conjunctive query:

$$Q(x) = \exists y\,(Scientist(x) \wedge isAuthorOf(x, y) \wedge Article(y) \wedge yearOfPublication(y, 2008)).$$

Semantic Web Knowledge Base

- Let **I** be the disjoint union of two sets **P** and **O** of Web pages and Web objects, respectively.
- A semantic annotation $\mathcal{A}_a$ for a Web page or object $a \in \mathbf{P} \cup \mathbf{O}$ is a finite set of concept membership axioms $A(a)$, role membership axioms $P(a, b)$, and attribute membership axioms $U(a, v)$, where $A \in \mathbf{A}$, $P \in \mathbf{R}_A$, $U \in \mathbf{R}_D$, $b \in \mathbf{I}$, and $v \in \mathbf{V}$.
- A Semantic Web knowledge base $KB = (\mathcal{T}, (\mathcal{A}_a)_{a \in \mathbf{P} \cup \mathbf{O}})$ consists of a TBox $\mathcal{T}$ and one semantic annotation $\mathcal{A}_a$ for every Web page and object $a \in \mathbf{P} \cup \mathbf{O}$.

Example

Semantic Web knowledge base $KB = (\mathcal{T}, (\mathcal{A}_a)_{a \in \mathbf{P} \cup \mathbf{O}})$:

- Set of individuals $\mathbf{I} = \mathbf{P} \cup \mathbf{O}$, where $\mathbf{P} = \{i_1\}$ is the set of Web pages, and $\mathbf{O} = \{i_2, i_3, i_4\}$ is the set of Web objects on i_1;
- TBox $\mathcal{T}$ as above;
- Semantic annotations $\mathcal{A}_a$:

$$
\begin{aligned}
\mathcal{A}_{i_1} &= \{contains(i_1, i_2), contains(i_1, i_3), contains(i_1, i_4)\},\\
\mathcal{A}_{i_2} &= \{PhDStudent(i_2),\ name(i_2, \text{“}mary\text{”}), isAuthorOf(i_2, i_3),\\
&\quad isAuthorOf(i_2, i_4)\},\\
\mathcal{A}_{i_3} &= \{ConferencePaper(i_3), title(i_3, \text{“}Semantic\ Web\ search\text{”})\},\\
\mathcal{A}_{i_4} &= \{JournalPaper(i_4),\ hasAuthor(i_4, i_2),\\
&\quad title(i_4, \text{“}Semantic\ Web\ search\ engines\text{”}),\\
&\quad yearOfPublication(i_4, 2008), keyword(i_4, \text{“RDF”})\}.
\end{aligned}
$$

Semantic Web Search Query (Syntax)

- An *atomic formula* (or *atom*) α is of one of the following forms:
  - $d(t)$, where d is an atomic datatype, and t is a term;
  - $A(t)$, where A is an atomic concept, and t is a term;
  - $P(t, t')$, where P is an atomic role, and t, t' are terms; and
  - $U(t, t')$, where U is an atomic attribute, and t, t' are terms.
- An *equality* has the form $=(t, t')$, where t and t' are terms.
- A *conjunctive formula* $\exists \mathbf{y}\, \phi(\mathbf{x}, \mathbf{y})$ is an existentially quantified conjunction of atoms α and equalities $=(t, t')$, which have free variables among $\mathbf{x}$ and $\mathbf{y}$.
- A Semantic Web search query $Q(\mathbf{x})$ is of form $\bigvee_{i=1}^{n} \exists \mathbf{y}_i\, \phi_i(\mathbf{x}, \mathbf{y}_i)$, where each ϕ_i with $i \in \{1, \ldots, n\}$ is a conjunction of atoms α (also called positive atoms), negated conjunctive formulas *not* ψ, and equalities $=(t, t')$, which have free variables among $\mathbf{x}$ and $\mathbf{y}_i$.

Example

Scientists who are either working for oxford university and did not receive their Ph.D. from that university, or who received their Ph.D. from oxford university but do not work for it:

$$Q_1(x) = (\mathit{Scientist}(x) \wedge \mathit{not}\ \mathit{doctoralDegree}(x, \text{``}\mathit{oxford\ university}\text{''}) \wedge \mathit{worksFor}(x, \text{``}\mathit{oxford\ university}\text{''})) \vee (\mathit{Scientist}(x) \wedge \mathit{doctoralDegree}(x, \text{``}\mathit{oxford\ university}\text{''}) \wedge \mathit{not}\ \mathit{worksFor}(x, \text{``}\mathit{oxford\ university}\text{''}));$$

Scientists of oxford university who are authors of at least one unpublished non-conference paper:

$$Q_2(x) = \exists y\, (\mathit{Scientist}(x) \wedge \mathit{worksFor}(x, \text{``}\mathit{oxford\ university}\text{''}) \wedge \mathit{isAuthorOf}(x, y) \wedge \mathit{not}\ \mathit{ConferencePaper}(y) \wedge \mathit{not}\ \exists z\, \mathit{yearOfPublication}(y, z)).$$

Semantic Web Search Query (Semantics)

Given a Semantic Web knowledge base *KB* and a positive (without negated conjunctive subqueries) Semantic Web search query $Q(\mathbf{x})$, an answer for $Q(\mathbf{x})$ to *KB* is a ground substitution θ for the variables $\mathbf{x}$ such that $KB \models Q(\mathbf{x}\theta)$.

An answer for a general $Q(\mathbf{x})$ to *KB* is a ground substitution θ for $\mathbf{x}$ such that $KB \models Q^{+}(\mathbf{x}\theta)$ and $KB \not\models Q^{-}(\mathbf{x}\theta)$, where:

$$Q^{+}(\mathbf{x}) = \bigvee_{i=1}^{n} \exists \mathbf{y}_i \, \phi_{i,1}(\mathbf{x}, \mathbf{y}_i) \wedge \cdots \wedge \phi_{i,l}(\mathbf{x}, \mathbf{y}_i) \text{ and}$$

$$Q^{-}(\mathbf{x}) = \bigvee_{i=1}^{n} \exists \mathbf{y}_i \, \phi_{i,1}(\mathbf{x}, \mathbf{y}_i) \wedge \cdots \wedge \phi_{i,l}(\mathbf{x}, \mathbf{y}_i) \wedge (\phi_{i,l+1}(\mathbf{x}, \mathbf{y}_i) \vee \cdots \vee \phi_{i,m}(\mathbf{x}, \mathbf{y}_i)) .$$

Example

$$Q(x) = \exists y\,(\mathit{Scientist}(x) \wedge \mathit{isAuthorOf}(x, y) \wedge \mathit{JournalPaper}(y) \wedge \exists z\, \mathit{yearOfPublication}(y, z)).$$

An answer for $Q(x)$ to KB is $\theta = \{x/i_2\}$. Recall that i_2 represents the scientist *mary*.

$$Q(x) = \exists y\,(\mathit{Article}(x) \wedge \mathit{hasAuthor}(x, y) \wedge \mathit{name}(y, \text{``}\mathit{mary}\text{''}) \wedge \mathit{not}\ \mathit{JournalPaper}(x) \wedge \mathit{not}\ \exists z\, \mathit{yearOfPublication}(x, z)).$$

An answer for $Q(x)$ to KB is $\theta = \{\mathbf{x}/i_3\}$. Recall that i_3 is an unpublished conference paper entitled “*Semantic Web search*”.

Query Processing

- Reduction of a query Q to standard Web search queries.
- The TBox $\mathcal{T}$ must be considered during standard Web search.
  - Compile $\mathcal{T}$ via offline ontology reasoning into the ABox $\mathcal{A}$ of KB, yielding a completed ABox $\mathcal{A}'$. Then, search by standard Web search queries depending on Q.
- An offline ontology reasoning step, where roughly all semantic annotations of Web pages / objects are completed by logically entailed membership axioms.
- An online reduction to standard Web search, where Q is transformed into a collection of standard Web search queries of which the answers are used to construct the answer for Q.

Offline Ontology Reasoning

Simple completion of KB is the Semantic Web knowledge base $KB' = (\emptyset, (\mathcal{A}_a{}')_{a \in \mathbf{P} \cup \mathbf{O}})$ such that every $\mathcal{A}_a{}'$ is the set of all $A(a)$, $P(a, b)$, and $U(a, v)$, where $A \in \mathbf{A}$, $P \in \mathbf{R}_A$, $U \in \mathbf{R}_D$, $b \in \mathbf{I}$, and $v \in \mathbf{V}$, that logically follow from $\mathcal{T} \cup \bigcup_{a \in \mathbf{P} \cup \mathbf{O}} \mathcal{A}_a$.

Evaluating SW search queries is correct but not complete (i.e., all answers are correct, but some answers may be missing).

Existentially quantified variables in the search query may refer to incompletely specified existentially quantified entries in the SW KB (not connected to concrete individuals and values).

Theorem. Let *KB* be a satisfiable SW KB over DL-Lite$_{\mathcal{A}}$. Let $Q(\mathbf{x})$ be a positive SWS query such that all existentially quantified variables occur only in safe positions, and let θ be a ground substitution for $\mathbf{x}$. Then, θ is an answer for $Q(\mathbf{x})$ to *KB* iff θ is an answer for $Q(\mathbf{x})$ to the simple completion of *KB*.

Corollary. Let *KB* be a satisfiable SW KB over DL-Lite$_{\mathcal{A}}$. Let $Q(\mathbf{x})$ be a (general) SWS query such that all existentially quantified variables occur only in safe positions, and let θ be a ground substitution for $\mathbf{x}$. Then, θ is an answer for $Q(\mathbf{x})$ to *KB* iff θ is an answer for $Q^{+}(\mathbf{x})$ but not an answer for $Q^{-}(\mathbf{x})$ to the simple completion of *KB*.

Special cases:

- query contains no existentially quantified variables;
- SW KB contains no existentially quantified variables in rule heads (i.e., is equivalent to a Datalog program).

Theorem. Given a SW KB *KB* over DL-Lite$_{\mathcal{A}}$, deciding (a) whether *KB* is satisfiable and (b) whether a given ground atom is in the simple completion of *KB* can both be done in polynomial time in general and in LOGSPACE in the size of the ABox of *KB* in the data complexity.

Theorem. Let *KB* be a SW KB over *DL-Lite*$_{\mathcal{A}}$. Then, (a) the size of the simple completion of *KB* is polynomial, and (b) computing it can be done in polynomial time, both in the size of *KB*.

Online Reduction to Web Search

Search queries where all free variables in negated conjunctive formulas and in equalities also occur in positive atoms are safe.

PositiveSemanticWebSearchQuery(Q)

$(Q^{x_1}, \ldots, Q^{x_n}) \leftarrow$ PositiveParse(Q);
For $i = 1$ To n Do $I_i \leftarrow$ WebSearchQuery(Q^{x_i});
For $i = 1$ To n Do $(R_j)_{j \in J_i} \leftarrow$ FillRelations(I_i);
Return $\pi_{\text{FREE}(Q)}(\bowtie_{i=1}^{n} \bowtie_{j \in J_i} R_j)$.

SemanticWebSearchQuery(Q)

$(Q_0, Q_1, \ldots, Q_m) \leftarrow$ Parse(Q);
For $i = 0$ To m Do
    $R_i \leftarrow$ PositiveSemanticWebSearchQuery(Q_i);
Return $\pi_{\text{FREE}(Q)}(\{t \in R_0 \mid \forall 1 \leqslant i \leqslant m \, \forall t_i \in R_i \colon t[R_i] \neq t_i\})$.

Ranking Answers

Generalization of PageRank: rather than considering only Web pages and the link structure between Web pages (expressed through the role *links_to* here), we also consider Web objects, which may occur on Web pages (expressed through the role *contains*), and which may also be related to other Web objects via other roles.

PageRank of a Web page or an object a:

$$R(a) = d \cdot \sum_{b \in B_a} R(b) \,/\, N_b + (1 - d) \cdot E(a)\,,$$

where (i) B_a is the set of all Web pages and Web objects that relate to a, (ii) N_b is the number of Web pages and Web objects that relate from b, (iii) d is a damping factor, and (iv) E associates with every Web page and every Web object a source of rank.

Implementation and Experiments

We have implemented a prototype for a semantic desktop search engine and obtained experimental results:

- the completed annotations are also rather small in practice;
- the online desktop search procedure scales quite well to very large collections of standard pages, annotation pages, and background ontologies;
- very high precision and recall; in many cases, SWS queries exactly describe the desired answer sets, resulting into a precision and a recall of 1.

Experiments: Size of Completed Annotations

Ontology	Average Size of a Completed Annotation (bytes)
FINITE-STATE-MACHINE[1]	202
SURFACE-WATER-MODEL[1]	173
NEW-TESTAMENT-NAMES[1]	229
SCIENCE[1]	146
FINANCIAL[2]	142

[1] From the Protégé Ontology Library:
http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library

[2] FINANCIAL ontology:
http://www.cs.put.poznan.pl/alawrynowicz/financial.owl

Experiments: Online Query Processing Time

10 queries $Q_1, \ldots, Q_{10}$ on a randomly generated SW KB (relative to the running SCIENCE ontology) with 5000 semantic annotations and 590270 facts:

(1) professors giving the course c_{12}:

$$Q_1(x) = Professor(x) \wedge teacherOf(x, c_{12})\,.$$

(2) professors giving the course c_{12} but not the course c_{20}:

$$Q_2(x) = Professor(x) \wedge teacherOf(x, c_{12}) \wedge not\ teacherOf(x, c_{20})\,.$$

(3) scientists working for o_{12} and authoring a_4, or scientists working for o_3 and authoring a_{25}:

$$Q_3(x) = (Scientist(x) \wedge worksFor(x, o_{12}) \wedge hasWritten(x, a_4)) \vee (Scientist(x) \wedge worksFor(x, o_3) \wedge hasWritten(x, a_{25}))\,.$$

(4) scientists working for u_{11} but not having a doctorate from u_{11}, or scientists having a doctorate from u_{11} but not working for u_{11}:

$$Q_4(x) = (Scientist(x) \wedge worksFor(x, u_{11}) \wedge not\ doctoralDegree(x, u_{11})) \vee (Scientist(x) \wedge doctoralDegree(x, u_{11}) \wedge not\ worksFor(x, u_{11}))\,.$$

Total time used (in ms) and number of returned URIs for processing the 10 queries $Q_1, \ldots, Q_{10}$ on the SW KB:

Query	Total Time (ms)		No. URIs
	FoIKS-2010 Prototype	New Prototype	
$Q_1(x)$	12123	**204**	613
$Q_2(x)$	5893	**27**	116
$Q_3(x)$	20858	**153**	582
$Q_4(x)$	14592	**91**	529
$Q_5(x)$	23001	**521**	679
$Q_6(x)$	16264	**220**	204
$Q_7(x)$	43847	**976**	687
$Q_8(x)$	4979	**10**	20
$Q_9(x)$	38971	**870**	687
$Q_{10}(x)$	54403	**884**	671

Total time used (in ms) by Corese and by our new prototype, along with the number of returned URIs, for processing 10 queries $Q_1, \ldots, Q_{10}$ on a randomly generated SW KB:

Query	Total Time (ms)		No. URIs
	Corese	New Prototype	
$Q_1(x)$	531	**115**	946
$Q_2(x)$	420	**43**	313
$Q_3(x)$	581	**226**	1942
$Q_4(x)$	395	**225**	1896
$Q_5(x)$	402	**76**	613
$Q_6(x)$	391	**45**	335
$Q_7(x)$	336	**4**	7
$Q_8(x)$	556	**209**	1252
$Q_9(x)$	521	**10**	32
$Q_{10}(x)$	557	**155**	970

Experiments: Precision and Recall

Precision and a recall of 10 Google Web search queries (compared to our Semantic Web search queries of precision and recall 1) addressed to the CIA World Fact Book:
http://www.cia.gov/library/publications/the-world-factbook/
relative to the WORLD-FACT-BOOK ontology:
http://www.ontoknowledge.org/oil/case-studies/

(1) countries having a common border with Austria:

$Q_1(x) = Country(x) \wedge borderCountries(x, Austria)$,
"border countries" Austria ;

(2) countries having Bulgaria as exports partners:

$Q_2(x) = Country(x) \wedge exportsPartners(x, Bulgaria)$,
$''$exports - partners$''$ Bulgaria ;

(3) countries in which Italian is spoken:

$Q_3(x) = Country(x) \wedge languages(x, Italian)$,
languages Italian ;

(4) countries importing tobacco:

$Q_4(x) = Country(x) \wedge importsCommodities(x, tobacco)$,
$''$imports - commodities$''$ tobacco ;

Precision and recall of Google vs. SW search:

Query	Results Google	Correct Results	Correct Results Google	**Precision Google**	**Recall Google**
$Q_1(x)$	17	8	8	0.47	1
$Q_2(x)$	19	5	5	0.26	1
$Q_3(x)$	21	13	13	0.62	1
$Q_4(x)$	51	10	10	0.2	1
$Q_5(x)$	24	4	4	0.17	1
$Q_6(x)$	229	253	229	1	0.91
$Q_7(x)$	33	32	32	0.97	1
$Q_8(x)$	11	13	11	1	0.85
$Q_9(x)$	45	7	7	0.16	1
$Q_{10}(x)$	6	3	1	0.17	0.33

Summary and Outlook

Summary:

- Semantic search on the Web, where standard Web pages are combined with background ontologies, on top of standard Web search engines and ontological inference technologies.
- Formal model behind this approach. Generalized PageRank technique. Technique for processing semantic search queries for the Web, consisting of an offline ontological inference step and an online reduction to standard Web search queries.
- Implementation in desktop search along with very promising experimental results.

Outlook:

- Real Web implementation.
- Transforming plain natural language search strings into the presented semantic search queries for the Web.
- Use of probabilistic ontologies?

References

- Bettina Fazzinga, Giorgio Gianforme, Georg Gottlob, and Thomas Lukasiewicz. Semantic Web search based on ontological conjunctive queries. In *Proceedings FoIKS 2010*, pp. 153–172. *LNCS* 5956, Springer, 2010.
- Bettina Fazzinga, Giorgio Gianforme, Georg Gottlob, and Thomas Lukasiewicz. Semantic Web search based on ontological conjunctive queries. Accepted for publication in *Journal of Web Semantics*.
- Claudia d'Amato, Nicola Fanizzi, Bettina Fazzinga, Georg Gottlob, and Thomas Lukasiewicz. Combining Semantic Web search with the power of inductive reasoning. In *Proceedings SUM 2010*, pp. 137–150. *LNCS* 6379, Springer, 2010.