Inductive Reasoning and Semantic Web Search

Claudia d'Amato
Dipartimento di Informatica
Università di Bari, Italy
claudia.damato@di.uniba.it

Floriana Esposito
Dipartimento di Informatica
Università di Bari, Italy
esposito@di.uniba.it

Nicola Fanizzi
Dipartimento di Informatica
Università di Bari, Italy
fanizzi@di.uniba.it

Bettina Fazzinga
DEIS, Università della
Calabria, Italy
bfazzinga@deis.unical.it

Georg Gottlob*
University of Oxford, UK
georg.gottlob@
comlab.ox.ac.uk

Thomas Lukasiewicz†
University of Oxford, UK
thomas.lukasiewicz@
comlab.ox.ac.uk

ABSTRACT

Extensive research activities are recently directed towards the Semantic Web as a future form of the Web. Consequently, Web search as the key technology of the Web is evolving towards some novel form of Semantic Web search. A very promising recent approach to such Semantic Web search is based on combining standard Web search with ontological background knowledge and using standard Web search engines as the main inference motor of Semantic Web search. In this paper, we propose to further enhance this approach to Semantic Web search by the use of inductive reasoning. This adds the important ability to handle inconsistencies, noise, and incompleteness, which often occur in distributed and heterogeneous environments, such as the Web. We report on a prototype implementation of the new approach and extensive experimental results.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*information search and retrieval*; H.2.4 [**Database Management**]: Systems—*query processing*

General Terms

Algorithms, Languages, Theory

Keywords

Inductive reasoning, Semantic Web search, semantic search, Semantic Web, conjunctive queries, ontologies, description logics

1. INTRODUCTION

Web search as the key technology of the Web is about to change radically with the development of the *Semantic Web* (SW) [1]. As a consequence, the elaboration of a new search technology for the Semantic Web [3], called *Semantic Web search*, is currently an extremely hot topic, both in Web-related companies and in academic research. The research can be roughly divided into two main directions: (1) the most common one is to develop a new form of search for searching the pieces of data and knowledge that are encoded in the new representation formalisms of the SW; (2) the less explored direction is to use the data and knowledge of the SW in order to add some semantics to Web search. A very promising recent representative of the second direction to SW search has been presented in [4]. The approach is based on (1) using ontological (unions of) conjunctive queries (which may contain negated subqueries) as SW search queries, (2) combining standard Web search with ontological background knowledge, (3) using the power of SW formalisms and technologies, and (4) using standard Web search engines as the main inference motor of SW search. It consists of an offline ontology compilation step, based on deductive reasoning techniques, and an online query processing step. In this paper and in [5], we propose to further enhance this approach by the use of inductive reasoning techniques for the offline ontology compilation step. To our knowledge, this is the first combination of SW search with inductive reasoning. The main contributions can be summarized as follows: **(1)** We develop a combination of SW search [4] with an inductive reasoning technique (based on similarity search [6] for retrieving the resources that likely belong to a query concept [2]). The latter serves in an offline ontology compilation step to compute completed semantic annotations. **(2)** Importantly, the new approach can handle inconsistencies, noise, and incompleteness in SW knowledge bases (KBs), which are all very likely to occur in distributed and heterogeneous environments, such as the Web. **(3)** We report on a prototype implementation and extensive experimental evaluation in the framework of desktop search.

2. SEMANTIC WEB SEARCH

Our proposed SW search system [4] consists of an *Interface*, a *Query Evaluator*, and an *Inference Engine*. The Query Evaluator is implemented on top of standard Web *Search Engines*. Standard *Web* pages and their objects are enriched by *Annotation* pages, based on an *Ontology*. We thus assume that there are semantic annotations to standard Web pages and to objects on standard Web pages. Such annotations are starting to be widely available for a large class of Web resources, especially with the Web 2.0. They may also be automatically learned from the Web pages and the ob-

*Computing Laboratory and Oxford-Man Institute of Quantitative Finance. G. Gottlob's work was supported by the EPSRC grant EP/E010865/1 "Schema Mappings and Automated Services for Data Integration and Exchange". G. Gottlob also gratefully acknowledges a Royal Society Wolfson Research Merit Award.

†Computing Laboratory. Alternative affiliation: Institut für Informationssysteme, TU Wien, Austria. T. Lukasiewicz's work was supported by the DFG under the Heisenberg Programme.

SAC'10 March 22-26, 2010, Sierre, Switzerland.

jects to be annotated, and/or they may be extracted from existing ontological KBs on the SW. Another assumption that we make is that Web pages and their objects have unique identifiers. For example, a Web page i_1 may contain information about a Ph.D. student i_2, called Mary, and two of her papers: a conference paper i_3 entitled "*Semantic Web search*" and a journal paper i_4 entitled "*Semantic Web search engines*" and published in 2008. There may now exist one semantic annotation each for the Web page, the Ph.D. student Mary, the journal paper, and the conference paper. The semantic annotations of i_1, i_2, i_3, and i_4 are formally expressed as the sets of axioms $\mathcal{A}_{i_1}$, $\mathcal{A}_{i_2}$, $\mathcal{A}_{i_3}$, and $\mathcal{A}_{i_4}$, respectively:

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

Using an ontology containing some background knowledge, these semantic annotations are then further enhanced in an offline ontology compilation step, where the **Inference Engine** adds all properties that can be deduced from the semantic annotations and the ontology. In [4], this is performed by a deductive such step. Here and in [5], we propose and explore the exploitation of inductive reasoning. For 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:

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

Using this knowledge, we can derive from the above annotations that papers i_3 and i_4 are also articles, and both authored by John. These resulting (*completed*) semantic annotations of (objects on) standard Web pages are published as HTML Web pages with pointers to the respective object pages, so that they (in addition to the standard Web pages) can be searched by standard search engines.

The **Query Evaluator** reduces each SW search query of the user in an online query processing 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*. As an example of a 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:

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

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 can both be submitted to a standard Web search engine, such as Google. The result of the original query Q is then built from the results of the two queries Q_1 and Q_2.

3. INDUCTIVE REASONING

We now illustrate the main advantages of using inductive rather than deductive reasoning in SW search, namely, that inductive reasoning (differently from deductive reasoning) can handle inconsistencies, noise, and incompleteness in SW knowledge bases.

Since inductive reasoning is based on the majority vote of the individuals in the neighborhood, it may be able to give a correct classification even in case of inconsistent knowledge bases. This aspect is illustrated by the following example.

Example 3.1 *Consider the description logic (DL) knowledge base $KB = (\mathcal{T}, \mathcal{A})$ that consists of the following TBox $\mathcal{T}$ and ABox $\mathcal{A}$:*

$$\mathcal{T} = \{\mathit{Man} \equiv \mathit{Male} \sqcap \mathit{Human};\ \mathit{Professor} \equiv \mathit{Person} \sqcap \exists \mathit{abilitatedTo}.\mathit{Teaching} \sqcap \exists \mathit{isSupervisorOf}.\mathit{PhDThesis} \sqcap \mathit{Researcher};\ \mathit{Researcher} \equiv \mathit{GraduatePerson} \sqcap \exists \mathit{worksFor}.\mathit{ResearchInstitute} \sqcap \neg\exists \mathit{isSupervisorOf}.\mathit{PhDThesis}; \ldots\};$$
$$\mathcal{A} = \{\mathit{Professor}(\mathit{Franz});\ \mathit{isSupervisorOf}(\mathit{Franz}, \mathit{DLThesis});\ \mathit{Professor}(\mathit{John});\ \mathit{isSupervisorOf}(\mathit{John}, \mathit{RoboticsThesis});\ \mathit{Professor}(\mathit{Flo});\ \mathit{isSupervisorOf}(\mathit{Flo}, \mathit{MLThesis});\ \mathit{Researcher}(\mathit{Nick});\ \mathit{Researcher}(\mathit{Ann});\ \mathit{isSupervisorOf}(\mathit{Nick}, \mathit{SWThesis}); \ldots\}.$$

Actually, Nick is a professor, indeed, he is the supervisor of a PhD thesis in $\mathcal{A}$. However, by mistake, he is asserted to be a researcher in $\mathcal{A}$, and by the axiom for Researcher in $\mathcal{T}$, he cannot be the supervisor of any PhD thesis. Hence, KB is inconsistent, and thus a deductive reasoner cannot answer whether Nick is a professor or not (since everything can be deduced from an inconsistent knowledge base). On the contrary, by inductive reasoning, it is highly probable that the returned classification result is that Nick is an instance of Professor, because the most similar individuals are Franz, John, and Flo, and all of them vote for the concept Professor.

Inductive reasoning may also be able to give a correct classification in the presence of noise in a knowledge base (containing, e.g., incorrect concept and/or role membership assertions), which is illustrated by the following example.

Example 3.2 *Consider the DL knowledge base $KB = (\mathcal{T}', \mathcal{A})$, where the ABox $\mathcal{A}$ is as in Example 3.1 and the TBox $\mathcal{T}'$ is obtained from the TBox $\mathcal{T}$ of Example 3.1 by replacing the axiom for Researcher by the following axiom:*

$$\mathit{Researcher} \equiv \mathit{GraduatePerson} \sqcap \exists \mathit{worksFor}.\mathit{ResearchInstitute}\,.$$

Again, Nick is actually a professor, but by mistake asserted to be a researcher in KB. But due to the slightly modified axiom for Researcher, there is no inconsistency in KB anymore. By deductive reasoning, however, Nick turns out to be a researcher, whereas by inductive reasoning, it is highly probable that the returned classification result is that Nick is an instance of Professor, as above.

Clearly, inductive reasoning may also be able to give a correct classification in the presence of incompleteness in a knowledge base. That is, inductive reasoning is not necessarily deductively valid, and may produce new knowledge.

Example 3.3 *Consider the DL knowledge base $KB = (\mathcal{T}', \mathcal{A}')$, where the TBox $\mathcal{T}'$ is as in Example 3.2 and the ABox $\mathcal{A}'$ is obtained from the ABox $\mathcal{A}$ of Example 3.1 by removing the axiom $\mathit{Researcher}(\mathit{Nick})$. Then, the resulting knowledge base is neither inconsistent nor noisy, but it is now incomplete. Nonetheless, by the same line of argumentation as above, it is highly probable that by inductive reasoning, Nick is an instance of Professor.*

4. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Sci. Am.*, 284:34–43, 2001.

[2] C. d'Amato, N. Fanizzi, and F. Esposito. Query answering and ontology population: An inductive approach. In *Proc. ESWC-2008*.

[3] L. Ding, T. W. Finin, A. Joshi, Y. Peng, R. Pan, and P. Reddivari. Search on the Semantic Web. *IEEE Computer*, 38(10):62–69, 2005.

[4] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. Semantic Web search based on ontological conjunctive queries. In *Proceedings FoIKS-2010. LNCS*, Springer, 2010.

[5] C. d'Amato, N. Fanizzi, B. Fazzinga, G. Gottlob, and T. Lukasiewicz. Combining Semantic Web search with the power of inductive reasoning. In *Proceedings URSW-2009*, pp. 15–26. *CEUR Workshop Proceedings* 527, CEUR-WS.org, 2009.

[6] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search — The Metric Space Approach*. Springer, 2006.