Ontology Mediated Information Extraction with MASTRO SYSTEM-T

Domenico Lembo[1], Yunyao Li[2], Lucian Popa[2], Kun Qian[2], and
Federico Maria Scafoglieri[⋆1]

[1] Dip. di Ingegneria Informatica, Automatica e Gestionale
Sapienza Università di Roma, Rome, Italy
[2] IBM Almaden Research Center, San Jose, California

Abstract. In several data-centric application domains, the need arises to extract valuable information from unstructured text documents. The recent paradigm of Ontology Mediated Information Extraction (OMIE) faces this problem by taking into account the knowledge expressed by a domain ontology, and reasoning over it to improve the quality of extracted data. MASTRO SYSTEM-T is a novel tool for OMIE, developed by Sapienza University and IBM Almaden Research. In this work, we demonstrate its usage for information extraction over real-world financial text documents from the U.S. EDGAR system.

Introduction

One of the basic problems of the data-centric information era is the processing of huge amount of unstructured data. If the information inside them is to be automatically manipulated and analyzed, it must be first rearranged into a structured form in which the relevant "facts" can be easily accessed.

Information Extraction (IE) provides support to this problem. It refers to the task of automatically organizing gathered data into a structured representation, typically a spread-sheet or a database [11, 6, 4]. Various statistical, rule-based, and learning based approaches for IE have been proposed along the years, leveraging techniques from NLP, machine learning, computational linguistics, databases and knowledge representation (see, e.g., [7, 2, 5, 1]). In this frame of reference, ontologies, which provide formal and explicit specifications of conceptualizations, have been recognized to play an important role in IE [12]. However, despite ontology-based IE has been so far the subject of several investigations [12, 10], how to exploit the reasoning abilities offered by an ontology to improve the extraction process has not yet been specifically studied.

Ontology Mediated Information Extraction (OMIE) [9, 8] is a new paradigm for IE which aims at filling this gap. It properly seeks to use the semantic knowledge expressed in ontologies to improve query answering over unstructured data (specifically raw text).

[⋆] Work done while at IBM Almaden - Research. Work supported in part by European Research Council under the European Union's Horizon 2020 Programme through the ERC Advanced Grant WhiteMech (No. 834228)

In this work, we demonstrate MASTRO SYSTEM-T, a new OMIE system born from a collaboration between the University of Rome “La Sapienza” and IBM Research Almaden. In particular, after a brief presentation of the system architecture and its main functionalities, we show an OMIE application involving a set of real-world financial text documents coming from the U.S repository of Electronic Data Gathering, Analysis and Retrieval system (EDGAR). Interestingly, with MASTRO SYSTEM-T we are able to extract data at query time, without having to materialize them in advance. We discuss this feature together with some preliminary experiments that show how ontology reasoning allow us to increase the quality of the extracted data.

System Overview

The OMIE framework, on which MASTRO SYSTEM-T is based, is an adaptation of the well-know framework of Ontology Based Data Access (OBDA) [13]. In an OBDA system, an ontology is mapped to an external source database through declarative mappings, which specify the semantic relationship between the ontology vocabulary and the data (mainly relational) at the sources. In OMIE the data source is instead a repository of unstructured text documents, which are “linked” to the ontology through so-called extraction assertions.

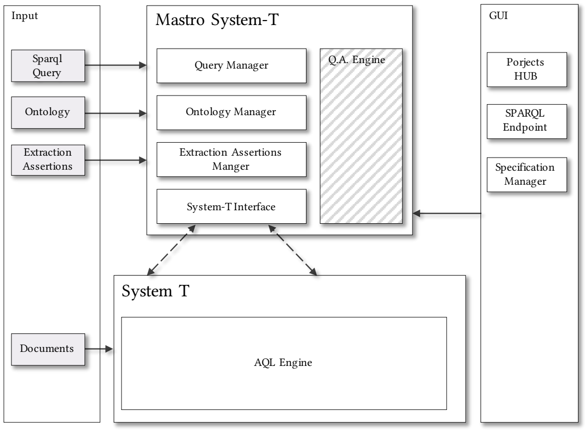


Fig. 1: MASTRO SYSTEM-T Architecture

This connection between OBDA and OMIE is also reflected in the implementation of our tool. MASTRO SYSTEM-T, whose architecture is showed in Fig. 1, is a specific tuning of the OBDA engine MASTRO [1] [3] in order to interface it with SYSTEM-T [1], an IE commercial tool developed at IBM Almaden. The inputs to the system are:

- An ontology, specified in any of the standard syntaxes for `OWL 2`. The ontology is automatically approximated by MASTRO in the standard profile OWL2QL, to guarantee tractability of query answering.

[1] `http://obdasystems.com/mastro`

- A set of extraction assertions (EAs) of the form $\phi(\vec{x}) \rightsquigarrow P(\vec{x})$, where P is a predicate of the ontology, $\phi(\vec{x})$ is a rule-based extractor, and $\vec{x}$ are “frontier variables”, through which, intuitively, data extracted from the source documents instantiate the ontology predicate P [9]. EAs are managed by the ‘Extraction Assertion Manager’ module. The extractors are specified into a declarative rule-based language, and can be combined together with relational algebra operators. Specifically, they are written in AQL, a concrete language used by SYSTEM-T, which is in charge of their processing. In simple terms, SYSTEM-T evaluates extractors over a text and produces a set of spans, i.e., pairs of indexes that identify substrings in the text that are used to construct the individuals that instantiate the ontology.
- A set of textual documents, which are managed by SYSTEM-T.
- The user’s queries, expressed in standard SPARQL, which are parsed and managed by the ‘Query Manager’ module.

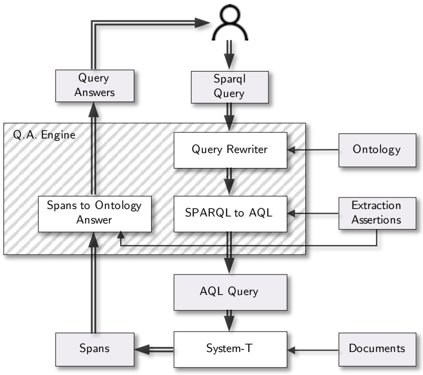
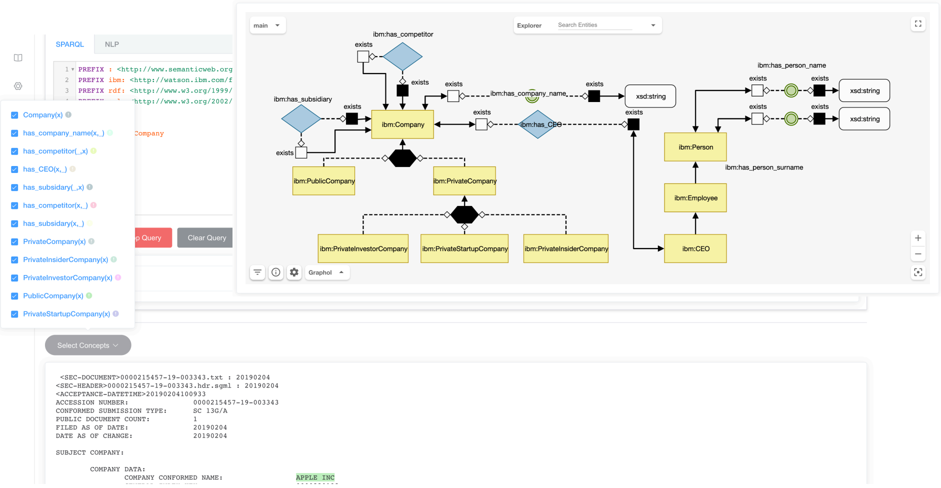


Fig. 2: Query Answering Workflow

Note that, following the principles of OBDA, in OMIE, the facts of the ontology are not materialized, but they are virtually defined through the extraction assertions.

The main reasoning service is Query Answering (QA), which is carried out through query rewriting techniques adapted from those used in MASTRO, as described in [9]. MASTRO SYSTEM-T computes answers to the user’s SPARQL queries posed over the ontology by transforming them into AQL extractors and delegating their execution for information extraction from a given document to SYSTEM-T. MASTRO SYSTEM-T triggers only the extraction assertions useful to generate the answers to the user’s query at hand and returns always the most up-to-date answer. This is particularly suited for dynamic scenarios, where source documents change frequently and query answers cannot be computed on the basis of outdated materializations.

In a nutshell, the query transformation process realized by the ‘QA Engine’ includes an ontology-based query rewriting phase, and a further reformulation step that uses extraction assertions to transform the query over the ontology into a set of extractors to be executed over the text documents. The complete workflow is illustrated in Fig. 2.



```
PREFIX : <http://www.semanticweb.org
PREFIX ibm: <http://watson.ibm.com/f
PREFIX rdf: <http://www.w3.org/1999/
            <http://www.w3.org/2002/

             Company
```

☑ Company(x)
☑ has_company_name(x,_)
☑ has_competitor(_,x)
☑ has_CEO(x,_)
☑ has_subsidiary(_,x)
☑ has_competitor(x,_)
☑ has_subsidiary(x,_)
☑ PrivateCompany(x)
☑ PrivateInsiderCompany(x)
☑ PrivateInvestorCompany(x)
☑ PublicCompany(x)
☑ PrivateStartupCompany(x)

```
 <SEC-DOCUMENT>0000215457-19-003343.txt : 20190204
<SEC-HEADER>0000215457-19-003343.hdr.sgml : 20190204
<ACCEPTANCE-DATETIME>20190204100933
ACCESSION NUMBER:                0000215457-19-003343
CONFORMED SUBMISSION TYPE:       SC 13G/A
PUBLIC DOCUMENT COUNT:           1
FILED AS OF DATE:                20190204
DATE AS OF CHANGE:               20190204

SUBJECT COMPANY:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 APPLE INC
```

Fig. 3: User Interface

Demonstration

We demonstrate MASTRO SYSTEM-T in a real world financial domain. The Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) is a public platform where companies acting in the U.S. are required by law to enter a range of information for government controls. EGARD is mainly composed by a large amount of raw text subject to significant updates over time. Since human effort is not sufficient to process this amount of data, there is the need for a mechanism that can automate the extraction phase by always providing the most up-to-date information and allowing data sharing and standardization. To prove the effectiveness of MASTRO SYSTEM-T in this context, we have created an ad-hoc ontology around the concept of company and a set of extraction assertions, and we have selected a set of text documents from EDGARD concerning the top five fortune companies. We then issued a set of queries, and processed them with and without the reasoning activated, in order to highlight its role in the extraction phase. To deactivate the reasoning we simply ask the system to skip the ontology-based query rewriting phase, which actually means that it ignores all ontology axioms. With respect to the tests that we have carried out, the reasoning mainly impacts on the recall. This is due to the fact that the compilation of the ontology inside the query leads to use a set of extractors that otherwise wouldn't have been triggered. As an example, in Table 1 we report the values of precision, recall and f-measure of the query that requires all companies, i.e., `SELECT ?X WHERE {?X a :Company}`.

Conclusions

Our preliminary tests show that reasoning over the ontology through MASTRO SYSTEM-T may improve the quality of certain extractions. We have also shown how

	Without Reasoning	With Reasoning	Gap
Precision	81.82%	82.71%	+0.89%
Recall	66.8%	76.26%	+9.46%
F-Measure	73.59%	79.35%	+5.76%

Table 1: Company query results

in our system data can be extracted at query-time, i.e., without having to materialize in advance all instances of the ontology, which always guarantees up-to-date answers. We are currently working to incorporate in MASTRO SYSTEM-T additional capabilities, e.g., to support entity linking, and reduce the design effort required for the specification of extraction assertions.

References

1. L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan. Systemt: An algebraic approach to declarative information extraction. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 128–137, 2010.
2. H. Cunningham. Gate, a general architecture for text engineering. *Comput. Humanit.*, 36(2):223–254, 2002.
3. G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, M. Ruzzi, and D. F. Savo. MASTRO: A reasoner for effective ontology-based data access. In *Proc. of the 1st Int. Workshop on OWL Reasoner Evaluation (ORE 2012)*, volume 858 of *CEUR*, 2012.
4. R. Fagin, B. Kimelfeld, F. Reiss, and S. Vansummeren. Document spanners: A formal approach to information extraction. *J. of the ACM*, 62(2):12, 2015.
5. R. Hoffmann, C. Zhang, X. Ling, L. S. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550, 2011.
6. D. Jurafsky and J. H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall, Pearson Education International, 2009.
7. J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th Int. Conf. on Machine Learning (ICML)*, pages 282–289, 2001.
8. D. Lembo, Y. Li, L. Popa, and F. M. Scafoglieri. Ontology mediated information extraction in financial domain with mastro system-t. In D. Burdick and J. Pujara, editors, *Proc. of the 6th Int. ACM Workshop on Data Science for Macro-Modeling, (DSMM 2020)*, pages 3:1–3:6. ACM, 2020.
9. D. Lembo and F. M. Scafoglieri. Ontology-based document spanning systems for information extraction. *Int. J. Semantic Comput.*, 14(1):3–26, 2020.
10. H. Saggion, A. Funk, D. Maynard, and K. Bontcheva. Ontology-based information extraction for business intelligence. In *Proc. of the 6th Int. Semantic Web Conference, and the 2nd Asian Semantic Web Conference (ISWC 2007 + ASWC 2007)*, pages 843–856, 2007.
11. S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.
12. D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Information Sciences*, 36(3):306–323, 2010.
13. G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence, (IJCAI 2018)*, pages 5511–5519, 2018.