Coupling Ontologies with Document Spanners

Domenico Lembo and Federico Maria Scafoglieri

DIAG, Sapienza Univeristà di Roma, Italy
{lembo,scafoglieri}@diag.uniroma1.it

Information extraction (IE) refers to the task of turning text documents into a structured form, in order to make the information contained therein automatically processable. Recently, Fagin et al. have carried out a foundational study on rule-based IE, and introduced a formal framework based on the notion of *(document) spanner* [6,7]. A *spanner* P is a function that maps a given string ($\mathbf{D}$) to a relation, i.e., a set of tuples, over its spans. A *span* is a pair of indices that identify substrings of $\mathbf{D}$. For example, given the string DL-2018_DL-2019, the spans $[4, 8\rangle$ and $[12, 16\rangle$ identify the substrings 2018 and 2019, respectively. Fagin et al. study possible representations of spanners and analyze how the use of some algebraic operations on the relations extracted from strings influences the spanner expressiveness. In particular, they consider spanners defined by regular expressions with capture variables (a.k.a. "*regex formulas*") and *relational algebra*. Regex formulas differ from classical regular expressions since they allow for mapping sub-matches of regular expressions, in the form of spans, to variables. For example, in the regex formula $x\{[0-9]^+\}$, x is a variable matching to spans of nonempty strings consisting only of digits. Applied, for instance, to the string DL-2018_DL-2019, it returns the (unary) relation $\{([4, 8\rangle), ([12, 16\rangle)\}$.

In our paper [8], we construct on the notion of document spanners and propose a formal framework for coupling them with ontologies. Through this framework we aim at structuring information extracted from text documents according to the intensional knowledge expressed by the ontology. To this aim, we adapt the well-known framework for Ontology-based Data Access (OBDA), in which an ontology is mapped to an external source database through declarative mappings, which specify the semantic relationship between the ontology vocabulary and the data at the sources [3,13]. OBDA is a powerful paradigm for data access and governance, for its ability to shift data management and integration at the conceptual level. In OBDA, however, ontologies have been used so far only on top of relational databases, with very few exceptions (as, e.g., [1]). In our paper we thus enrich OBDA with the capability of accessing unstructured information contained in text documents.

Our first contribution is the notion of Ontology-based document spanning (OBDS) system, in which an ontology is linked to text documents through *extraction assertions*, which in OBDS act exactly as mapping assertions in OBDA. In particular, an extraction assertion associates a document spanner P to a query q specified over the ontology. In OBDSs queries over the ontology are conjunctive queries (CQs), and document spanners are defined as regex formulas extended with the relational algebra operators union, projection, join and string selection (the class of such spanners is studied in [6] and is denoted with $[\![\mathrm{RGX}^{\{\cup,\pi,\bowtie,\zeta^=\}}]\!]$).

An OBDS $\mathcal{E}$ is thus a pair $\langle \mathcal{T}, \mathcal{R} \rangle$, where $\mathcal{T}$ is a Description Logic (DL) TBox and $\mathcal{R}$ is a set of extraction assertions of the form

$$P(v_1, \ldots, v_n) \rightsquigarrow \Psi(v_1, \ldots, v_n)$$

where $P(v_1, \ldots, v_n)$ is a document spanner in $[\![\mathrm{RGX}^{\{\cup,\pi,\bowtie,\zeta^=\}}]\!]$, associated to variables $v_1, \ldots, v_n$, and $\Psi(v_1, \ldots, v_n)$ is a CQ with free variables $v_1, \ldots, v_n$. Atoms of this CQ are built over $v_1, \ldots, v_n$, and possibly over other existential variables and/or constants, and *object terms* denoting individuals "constructed" from the spans returned by P when applied to a document (as in OBDA mappings [10]). We notice that extraction assertions we have defined correspond to a powerful form of GLAV mapping assertions [9,5]. We also say that assertions are GAV, if $\Psi(v_1, \ldots, v_n)$ does not contain existential variables. An interpretation $\mathcal{I}$ is a *model* for $\mathcal{E}$ w.r.t. a document $\mathbf{D}$ if $\mathcal{I}$ is a model for $\mathcal{T}$, and $\mathcal{I}$ satisfies $\mathcal{R}$ w.r.t. $\mathbf{D}$. As in OBDA, we adopt a sound interpretation for extraction assertions, according to which, intuitively, satisfying $\mathcal{R}$ w.r.t. $\mathbf{D}$ means that, for each extraction assertion in $\mathcal{R}$, $\Psi(\boldsymbol{t})$ evaluates to true in $\mathcal{I}$, for each tuple $\boldsymbol{t} = t_1, \ldots, t_n$ of substrings corresponding to the spans returned by P, where $\Psi(\boldsymbol{t})$ is the CQ in which every v_i is substituted with t_i (see [8] for details).

A second contribution of our paper is on query answering, i.e., how to compute the *certain answers* to queries posed over the ontology of an OBDS system $\mathcal{E} = \langle \mathcal{T}, \mathcal{R} \rangle$, i.e., those answers that hold in each model of $\mathcal{E}$. We focus on CQs and OBDS systems whose ontology is specified in the Description Logic (DL) *DL-Lite*$_R$ [2]. It is well-known that in OBDA, when the ontology is in *DL-Lite*$_R$ and mapping assertions are GLAV, CQ answering is first-order rewritable, i.e., it can be reduced to the evaluation of a first-order query over the underlying database [4]. Therefore, the rewriting has the same expressiveness of the database queries used in the mapping. A natural question is now whether a similar behaviour shows up also in *DL-Lite*$_R$ OBDS systems, i.e., whether we can reduce query answering to the execution of a document spanner of the same expressiveness of the spanners in $\mathcal{R}$. We positively answer to this question, by providing an algorithm that rewrites every CQ issued over a *DL-Lite*$_R$ OBDS system into a spanner belonging to $[\![\mathrm{RGX}^{\{\cup,\pi,\bowtie,\zeta^=\}}]\!]$. To this aim, we adapt to OBDS a technique from OBDA [4]. In a nutshell: we first split each extraction assertion into a pair of GAV and LAV assertions; we then rewrite the query according to the TBox by using, e.g., the PerfectRef algorithm [2]; we then further rewrite the query according to the "LAV part" of $\mathcal{R}$, by means of an algorithm for view-based query rewriting, like Minicon [11]; finally we rewrite the query with respect to the "GAV part" of $\mathcal{R}$, essentially by unfolding the query predicates with the spanners they are associated with. Since evaluating such spanners is polynomial in the size of the input document, we can also conclude that CQ answering in this setting is polynomial in data complexity.

We finally note that there is a lot of previous work on use of ontologies in IE (see, e.g., [12] for a survey). To the best of our knowledge, our paper is the first one studying query answering over ontologies populated from text documents, in the spirit of OBDA. Also, we believe that our OBDS framework may pave the way for an in-depth investigation of the role of ontologies in IE.

References

1. E. Botoeva, D. Calvanese, B. Cogrel, M. Rezk, and G. Xiao. OBDA beyond relational DBs: A study for MongoDB. In *Proc. of the 29th Int. Workshop on Description Logic (DL)*, 2016.
2. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning*, 39(3):385–429, 2007.
3. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Ontology-based data access and integration. In *Encyclopedia of Database Systems, Second Edition*. Springer, 2018.
4. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, and M. Ruzzi. Using OWL in data integration. In *Semantic Web Information Management - A Model-Based Perspective*, pages 397–424. Springer, 2009.
5. A. Doan, A. Y. Halevy, and Z. G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.
6. 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.
7. R. Fagin, B. Kimelfeld, F. Reiss, and S. Vansummeren. Declarative cleaning of inconsistencies in information extraction. *ACM Trans. on Database Systems*, 41(1):6:1–6:44, 2016.
8. D. Lembo and F. Scafoglieri. A formal framework for coupling document spanners with ontologies. In *Proc. of the 2nd IEEE Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2019.
9. M. Lenzerini. Data integration: A theoretical perspective. In *Proc. of the 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS)*, pages 233–246, 2002.
10. A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
11. R. Pottinger and A. Halevy. Minicon: A scalable algorithm for answering queries using views. *The VLDB Journal*, 10(2-3):182–198, 2001.
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)*, pages 5511–5519, 2018.