A Logic for Document Spanners

Dominik D. Freydenberger

Loughborough University, Loughborough, United Kingdom

Abstract

Document spanners are a formal framework for information extraction that was introduced by Fagin, Kimelfeld, Reiss, and Vansummeren (PODS 2013, JACM 2015). One of the central models in this framework are core spanners, which are based on regular expressions with variables that are then extended with an algebra. As shown by Freydenberger and Holldack (ICDT 2016), there is a connection between core spanners and $\mathsf{EC}^{\mathsf{reg}}$, the existential theory of concatenation with regular constraints. The present paper further develops this connection by defining SpLog, a fragment of $\mathsf{EC}^{\mathsf{reg}}$ that has the same expressive power as core spanners. This equivalence extends beyond equivalence of expressive power, as we show the existence of polynomial time conversions between this fragment and core spanners. This even holds for variants of core spanners that are based on automata instead of regular expressions. Applications of this approach include an alternative way of defining relations for spanners, insights into the relative succinctness of various classes of spanner representations, and a pumping lemma for core spanners.

This talk is based on the paper [5], which was presented at ICDT 2017.

1 Introduction

Fagin, Kimelfeld, Reiss, and Vansummeren [4] introduced *document spanners* as a formal framework for information extraction. Document spanners formalize the query language AQL that is used in IBM's SystemT. On an intuitive level, document spanners can be understood as a generalized form of searching in a text w: In its basic form, search can be understood as taking a search term u (or a regular expression α) and a word w, and computing all intervals of positions of w that contain u (or a word from $\mathcal{L}(\alpha)$). These intervals are called *spans*. Spanners generalize searching by computing *relations over spans* of w.

In order to define spanners, [4] introduced *regex formulas*, which are regular expressions with variables. Each variable x is connected to a subexpression α, and when α matches a subword of w, the corresponding span is stored in x (this behaves like the capture groups that are often used in real world implementation of search-and-replace functionality). *Core spanners* combine these regex formulas with the algebraic operators projection, union, join (on spans), and string equality selection.

Freydenberger and Holldack [6] connected core spanners to $\mathsf{EC}^{\mathsf{reg}}$, the existential theory of concatenation with regular constraints. Described very informally, $\mathsf{EC}^{\mathsf{reg}}$ is a logic that combines equations on words (like $x\mathtt{ab}y = y\mathtt{ba}x$) with positive logical connectives, and regular languages that constrain variable replacement. In particular, [6] showed that every core spanner can be converted into an $\mathsf{EC}^{\mathsf{reg}}$-formula, which can then be used to decide satisfiability. Furthermore, while every $\mathsf{EC}^{\mathsf{reg}}$-formula can be converted into an equisatisfiable core spanner, the resulting spanner cannot be used to evaluate the formula (as, due to details of the encoding, the input word w of the spanner needs to encode the formula).

This paper further explores the connection of core spanners and $\mathsf{EC}^{\mathsf{reg}}$. As main conceptual contribution, we introduce SpLog (short for *spanner logic*), a natural fragment of $\mathsf{EC}^{\mathsf{reg}}$ that has the same expressive power as core spanners. In contrast to the PSPACE-complete combined complexity of $\mathsf{EC}^{\mathsf{reg}}$-evaluation, the combined complexity of SpLog-evaluation is NP-complete, and its data complexity is in NL. As main technical result, we demonstrate

the existence of polynomial time conversions between SpLog and spanner representations (in both directions), even if the spanners are defined with automata instead of regex formulas.

As a consequence, SpLog can augment (or even replace) the use of regex formulas or automata in the definition of core spanners. Moreover, this shows that the PSPACE upper bounds from [6] for deciding satisfiability and hierarchicality of regex formula based spanners apply to automata based spanners as well. In addition to this, we adapt a pumping lemma for word equations to SpLog (and, hence, to core spanners). The main result also provides insights into the relative succinctness of classes of automata based spanners: While there are exponential trade-offs between various classes of automata, these differences disappear when adding the algebraic operators.

From a more general point of view, this paper can also be seen as an attempt to connect spanners to the research on equations on words and on groups (cf. Diekert [3, 2] for surveys), where $\mathsf{EC}^{\mathsf{reg}}$ has been studied as a natural extension of word equations. We shall see that SpLog is a natural fragment of $\mathsf{EC}^{\mathsf{reg}}$: On an informal level, SpLog has to express relations on a word w without using additional working space (which explains the friendlier complexity of evaluation, in comparison to $\mathsf{EC}^{\mathsf{reg}}$). This gives us reason to expect that SpLog can be applied to other models, like graph databases (as a related example of an application of $\mathsf{EC}^{\mathsf{reg}}$ for graph databases, Barceló and Muñoz [1] use a restricted class of $\mathsf{EC}^{\mathsf{reg}}$-formulas for which data complexity is also in NL).

References

1 P. Barceló and P. Muñoz. Graph logics with rational relations: the role of word combinatorics. In *Proc. CSL-LICS 2014*, 2014.

2 V. Diekert. Makanin's Algorithm. In M. Lothaire, editor, *Algebraic Combinatorics on Words*, chapter 12. Cambridge University Press, 2002.

3 V. Diekert. More than 1700 years of word equations. In *Proc. CAI 2015*, 2015.

4 R. Fagin, B. Kimelfeld, F. Reiss, and S. Vansummeren. Document spanners: A formal approach to information extraction. *J. ACM*, 62(2):12, 2015.

5 D. D. Freydenberger. A logic for document spanners. In *Proc. ICDT 2017*, 2017.

6 Dominik D. Freydenberger and Mario Holldack. Document spanners: From expressive power to decision problems. *Theor. Comput. Sys.*, 2017.