Efficient evaluation of regular patterns for information extraction

Antoine Amarilli Pierre Bourhis Stefan Mengel

Many approaches have been proposed for the task of *information extraction*. One recent such approach is to define the patterns to extract with extended regular expressions called *document spanners*. This has been explored in particular in the IBM SystemT project, which motivated many research works in database theory [JACM2015], [PODS2014]. Spanners have also been used for information extraction from CSV files, a common textual representation of tabular data [VLDB2016]. However, while this declarative approach makes it easier to describe extraction rules, the evaluation of these rules remains complex, in particular because the number of extracted occurrences can be very large.

To address this issue, one recent proposal is to build a compact in-memory representation of the pattern occurrences in the input document, and then to enumerate them one after the other using this structure. This approach, called *enumeration*, has received much attention recently, in particular for information extraction using spanners [PODS2018a], [PODS2018b], but also in the more classical context of queries on words expressed in monadic second-order logic [PODS2018c].

Sadly, these theoretical works do not immediately lead to practical algorithms for information extraction, because they usually assume that the spanners have been translated to deterministic automata, which generally causes an exponential blowup. The goal of this project is to develop efficient algorithms and implement them to perform practical information extraction from text. To do this, one direction is to adapt the methods proposed in our recent work [ICALP2017] and [ICDT2019].

A second goal of the project is to update the compact representation of the answers when the input text is modified. This problem has been studied in particular in [PODS2018c]. The challenge is again to propose algorithms that can scale efficiently in the input document.

Practical details. The internship will be co-advised by Antoine Amarilli (Télécom ParisTech), Pierre Bourhis (CNRS CRIStAL & Inria Lille), and Stefan Mengel (CNRS CRIL). It can take place at Télécom ParisTech in Paris, at INRIA Lille, or at CNRS Cril in Lens, to be discussed with the prospective student.

References

— [ICALP2017] : Antoine Amarilli, Pierre Bourhis, Louis Jachiet, Stefan Mengel. *A Circuit-Based Approach to Efficient Enumeration*. ICALP 2017.

— [ICDT2019] : Antoine Amarilli, Pierre Bourhis, Stefan Mengel, Matthias Niewerth. *Constant-Delay Enumeration for Nondeterministic Document Spanners*. ICDT 2019.

— [JACM2015] : Ronald Fagin, Benny Kimelfeld, Frederick Reiss, Stijn Vansummeren. *Spanners: A Formal Framework for Information Extraction.* JACM 2015.

— [PODS2014] : Benny Kimelfeld. *Database Principles in Information Extraction*. PODS 2014.

— [PODS2018a] : Dominik D. Freydenberger, Benny Kimelfeld, Liat Peterfreund. *Joining Extractions of Regular Expressions*. PODS 2018.

— [PODS2018b] : Fernando Florenzano, Cristian Riveros, Martín Ugarte, Stijn Vansummeren, Domagoj Vrgoc. *Constant Delay Algorithms for Regular Document Spanners*. PODS 2018.

— [PODS2018c] : Matthias Niewerth, Luc Segoufin. *Enumeration of MSO Queries on Strings with Constant Delay and Logarithmic Updates*. PODS 2018.

— [VLDB2016] : Marcelo Arenas, Francisco Maturana, Cristian Riveros, Domagoj Vrgoc. *A Framework for Annotating CSV-like Data*. VLDB 2016.