Abstract

SeqBIM

Efficient enumeration of regex matches

Antoine Amarilli

LTCI, Télécom Paris, Institut Polytechnique de Paris

Abstract

Database theory research has recently studied the task of *declarative information extraction*, i.e., extracting structured information by searching for relevant patterns in unstructured textual documents. This was initially motivated by IBM's SystemT tool [1]. The task is formalized using so-called *regex-formulas*, which are regular expressions with capture variables. For instance, if we have text containing names and email addresses of the form "`John Doe <john.doe@example.com>`", the following regex formula applied to the text will extract all pairs (x, y) of a name (mapped to variable x) and its corresponding email address (mapped to variable y):

$$x\{\texttt{[A-Z][a-z]* [A-Z][a-z]*}\} \texttt{<}y\{\texttt{[a-z.]+@[a-z.]+}\}\texttt{>}$$

Recent research on this topic has studied how to efficiently perform this extraction task: given a regex formula ϕ with variables X and a textual document D, find all possible assignments of the variables X to spans (i.e., intervals of positions) of D such that ϕ is satisfied. As the number of results may be huge, we are looking for an *enumeration algorithm* [2], which produces results one after the other in an anytime fashion. Specifically, we measure the *preprocessing* time before the first answer is found, and then measure the *delay* between any two successive answers.

The proposed talk will present this research area, recent results, and ongoing research directions. It will focus on our recent work with Pierre Bourhis, Stefan Mengel, and Matthias Niewerth, published at ICDT'19 [3] and distinguished at SIGMOD Research Highlights[1]. In this work, we showed that the matches of a regex-formula in a textual document can be enumerated with linear preprocessing in the input document and constant delay between each answer, with the complexity in the regex-formula being polynomial. Our work is supported by an implementation[2] which is benchmarked in our upcoming journal article [4].

The goal is to explore how these methods could relate to those of the SeqBIM community and identify use cases, e.g., efficiently locating patterns in genomic data.

References

[1] IBM Research. SystemT, 2018.

[2] Kunihiro Wasa. Enumeration of enumeration algorithms. *CoRR*, abs/1605.05102, 2016.

[3] Antoine Amarilli, Pierre Bourhis, Stefan Mengel, and Matthias Niewerth. Constant-delay enumeration for nondeterministic document spanners. In *ICDT*, 2019.

[4] Antoine Amarilli, Pierre Bourhis, Stefan Mengel, and Matthias Niewerth. Constant-delay enumeration for nondeterministic document spanners. Under review, 2020.

[1] https://sigmodrecord.org/2020/07/31/constant-delay-enumeration-for-nondeterministic-document-spanners/

[2] https://github.com/PoDMR/enum-spanner-rs