Internship Proposal

Enumerating Regular Language Matches on Dynamic Data

Antoine Amarilli <antoine.amarilli@telecom-paris.fr>

Background. This internship studies the problem of finding the matches of a regular language query on a large textual document. The simplest kind of such queries are *factor queries*: fixing a finite automaton A, and given a string w, determine all pairs of endpoints i, j such that the factor $w[i, j]$ is accepted by A. *Document spanners* [7] generalize these queries: they are finite automata with capture variables, and their matches are the assignments of the capture variables on w that makes the automaton accept. The motivation for document spanners comes from *information extraction*, i.e., automatically finding interesting patterns in large documents: document spanners are a declarative formalism to specify which data should be extracted.

The results of such an extraction task can be large, i.e., polynomial in the input document. Hence, the running time of an algorithm to solve this problem will be dominated in the worse case by the time needed to write down the result. For this reason, we study such tasks with a finer complexity measure, using the framework of *enumeration algorithms* [1]. In this framework, an algorithm must produce the query results in streaming, and we try to minimize the *preprocessing time* before the first solution is produced, and the subsequent *delay* between two consecutive solutions.

The problem of finding the matches of a fixed document spanner on a textual document is known to be solvable with *linear preprocessing* of the input document, and *constant delay* between two successive answers. This is known since 2006 [4], but was recently reproved using new techniques [2].

The goal of this internship is to study how to extend these results to *dynamic data*, when the underlying textual document can receive modifications. In this context, whenever the document is modified, we want to adjust the enumeration structure without re-running the preprocessing from scratch. Incremental enumeration problems of this sort have already been studied [5, 8, 9, 1]. The results of [9] imply that we can enumerate the results of a document spanner while supporting edits on the underlying document in logarithmic time.

In a different line of work, we have studied how to maintain simpler queries over dynamic words. In this *dynamic membership* problem, we fix a finite automaton A, we are given an input word, and we must handle substitution updates on the input word while maintaining the information of whether the entire word is accepted by A. Our work [3] shows that updates can be handled faster than logarithmic time for some regular languages. However, we do not know yet if these results can be extended to an enumeration algorithm for factor queries or document spanners.

Research topic: The goal of this internship is to combine the results on enumeration for dynamic words with our results on the dynamic membership problem, and study for which document spanners we can achieve linear-time preprocessing and constant-delay enumeration with a more efficient support for updates. For instance, one first task it to identify classes of factor queries for which updates can be supported in constant time as in [6].

[1] https://en.wikipedia.org/wiki/Enumeration_algorithm

Supervision and environment. The internship will take place in the DIG team of Télécom Paris and will be supervised by Antoine Amarilli, associate professor in the team. If necessary, the intern will also work with Louis Jachiet (Télécom Paris) and Charles Paperman (Université de Lille), along with Luc Segoufin (École normale supérieure). The internship is proposed in the context of the ANR binational project EQUUS[2]

References

[1] A. Amarilli, P. Bourhis, S. Mengel, and M. Niewerth. Enumeration on trees with tractable combined complexity and efficient updates. In *PODS*, 2019.

[2] A. Amarilli, P. Bourhis, S. Mengel, and M. Niewerth. Constant-delay enumeration for nondeterministic document spanners. *TODS*, 2020.

[3] A. Amarilli, L. Jachiet, and C. Paperman. Dynamic membership for regular languages. In *ICALP*, 2021.

[4] G. Bagan. MSO queries on tree decomposable structures are computable with linear delay. In *CSL*, 2006.

[5] A. Balmin, Y. Papakonstantinou, and V. Vianu. Incremental validation of XML documents. *TODS*, 29(4), 2004.

[6] C. Berkholz, J. Keppeler, and N. Schweikardt. Answering FO+MOD queries under updates on bounded degree databases. In *ICDT*, 2017.

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

[8] K. Losemann and W. Martens. MSO queries on trees: Enumerating answers under updates. In *CSL-LICS*, 2014.

[9] M. Niewerth and L. Segoufin. Enumeration of MSO queries on strings with constant delay and logarithmic updates. In *PODS*, 2018.

[2] `https://anr.fr/Project-ANR-19-CE48-0019`.