Évaluation efficace de motifs réguliers pour l'extraction de données

Antoine Amarilli Pierre Bourhis Stefan Mengel

L'extraction de données est un problème récurrent qui a été étudié de différentes manières. Une approche moderne consiste à définir ce que l'on souhaite extraire en utilisant des expressions régulières étendues appelées *spanners*. C'est en particulier le cas du projet SystemT d'IBM, qui a motivé de nombreux travaux de recherche en théorie de bases de données [JACM2015], [PODS2014]. Les spanners sont également utilisés pour l'extraction d'information dans les fichiers CSV, qui est le format le plus utilisé pour représenter des données tabulaires [VLDB2016] Cependant, si ces approches déclaratives facilitent la description des règles d'extraction, leur évaluation demeure complexe, notamment car le nombre de réponses est susceptible d'être très grand.

Pour remédier à cela, une approche récente consiste à construire une représentation compacte en mémoire des occurrences du motif sur le texte d'entrée, puis à les énumérer une par une à partir de cette structure. Cette approche, appelée *énumération*, a été très étudiée ces dernières années pour différents types de données, en particulier pour l'extraction d'information à l'aide de spanners [PODS2018a], [PODS2018b], mais également dans le cadre classique des requêtes sur les mots exprimée en logique monadique du second ordre [PODS2018c].

Hélas, ces travaux actuels ne permettent guère d'évaluer efficacement des motifs en pratique, car ils nécessitent de traduire les spanners vers des automates déterministes, ce qui engendre une explosion combinatoire. L'objectif de ce projet est de concevoir des algorithmes permettant une implémentation efficace de ces techniques pour l'extraction de données dans un texte. Pour cela, nous adapterons les méthodes développées pour énumérer les solutions de circuits proposées dans [ICALP2017] [ICDT2019] [ARXIV2018].

Dans un deuxième temps, nous nous intéresserons à comment on peut maintenir à jour la représentation compacte des réponses lorsque le texte d'entrée est modifiée. Ce problème a été étudié en particulier dans [PODS2018c]. Notre approche est là encore de proposer des algorithmes implémentables et passant à l'échelle.

Supervision et environnement

Ce stage sera encadré par Antoine Amarilli (Télécom ParisTech), conjointement avec Pierre Bourhis (CNRS CRiSTAL) et Stefan Mengel (CNRS CRIL). Il sera situé à Télécom ParisTech.

Bibliographie

— [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.

— [ARXIV2018] : Antoine Amarilli, Pierre Bourhis, Stefan Mengel, Matthias Niewerth. *Enumeration on Trees with Tractable Combined Complexity and Efficient Updates*. Under review, 2018.

— [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.