Updatable Extracted Views

Besat Kassaie*

David R. Cheriton School of Computer Science
University of Waterloo
bkassaie@uwaterloo.ca

Abstract

By identifying strings of interest in unstructured or semi-structured data sources, extractors provide selected data to populate relational records. In this work, we characterize extraction algorithms that are resilient to changes in their source documents intended to reflect predetermined changes to the extracted table, i.e., they produce updatable extracted views. We introduce and formalize the notion of *stable* information extraction algorithms and propose statically verifiable properties for such extractors. We further propose a translation mechanism to reflect updates on the views to updates on the input document. We propose a verification process for the stability of programs written in a significant subset of AQL, a commonly used declarative rule-based extraction language.

1 Introduction

When extracted relations or source documents are updated, we wish to ensure that those changes are propagated correctly. That is, we recommend that extracted relations be treated as materialized views over the document database. In this context, maintaining system consistency requires translating updates made to extracted views into corresponding updates in the original documents [2].

In this talk, I begin by exploring update-aware information extraction, highlighting the critical challenges that arise when dealing with updates. I then delve into our research on updatable extracted views, emphasizing their importance in the context of unstructured data cleaning. Additionally, I discuss the solutions we've developed for SystemT [3].

2 Unstructured Data Cleaning

Existing data cleaning techniques for unstructured data are often insufficient, as they are typically embedded within other processing steps, leaving the underlying data sources uncleaned [1]. Therefore, a cleaning process must be initiated from scratch for each new application.

In this work, we propose instead to clean documents and use the cleaned version in applications. To this end, we recommend to use document-at-a-time information extractors to create a relational view over documents. With an adequate extractor, data quality failures in documents can be revealed in extracted views, cleaning can then be applied over the extracted items, and finally the cleaned items can be transferred back into the source documents (or a copy if the original is to be preserved). To this end, we consider a document "clean" if its corresponding extracted view is clean. Within this context, we formally define stable extraction programs and show that, if an extraction program is stable, we can modify the document so that the synthetic version of an extracted view can be directly obtained from the altered text using the same extraction program. Additionally, we propose a verifier that evaluates core AQL programs for sufficient conditions to determine their stability.

References

[1] Priya Deshpande, Alexander Rasin, Roselyne Tchoua, Jacob D. Furst, Daniela Raicu, and Sameer K. Antani. Enhancing recall using data cleaning for biomedical big data. In *33rd IEEE CBMS 2020, Rochester, MN, USA, July 28-30, 2020*, pages 265–270. IEEE, 2020.

[2] Besat Kassaie and Frank Wm. Tompa. Predictable and consistent information extraction. In *Proc. DocEng '19: ACM*, pages 14:1–14:10. ACM, 2019.

[3] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan. An algebraic approach to rule-based information extraction. In *Proc. 24th ICDE*, pages 933–942. IEEE Computer Society, 2008.

*This talk is based on collaborative research with Frank Wm. Tompa: fwtompa@uwaterloo.ca.