Doleschal, Johannes; Kimelfeld, Benny; Martens, Wim; Peterfreund, Liat
Weight annotation in information extraction. (English) [Zbl 07471710]

Summary: The framework of document spanners abstracts the task of information extraction from text as a function that maps every document (a string) into a relation over the document’s spans (intervals identified by their start and end indices). For instance, the regular spanners are the closure under the Relational Algebra (RA) of the regular expressions with capture variables, and the expressive power of the regular spanners is precisely captured by the class of VSet-automata – a restricted class of transducers that mark the endpoints of selected spans.

In this work, we embark on the investigation of document spanners that can annotate extractions with auxiliary information such as confidence, support, and confidentiality measures. To this end, we adopt the abstraction of provenance semirings by Green et al., where tuples of a relation are annotated with the elements of a commutative semiring, and where the annotation propagates through the positive RA operators via the semiring operators. Hence, the proposed spanner extension, referred to as an annotator, maps every string into an annotated relation over the spans. As a specific instantiation, we explore weighted VSet-automata that, similarly to weighted automata and transducers, attach semiring elements to transitions. We investigate key aspects of expressiveness, such as the closure under the positive RA, and key aspects of computational complexity, such as the enumeration of annotated answers and their ranked enumeration in the case of ordered semirings. For a number of these problems, fundamental properties of the underlying semiring, such as positivity, are crucial for establishing tractability.

MSC:
03B70 Logic in computer science
68-XX Computer science

Keywords:
information extraction; regular document spanners; weighted automata; provenance semirings; K-relations

Software:
ProbLog

Full Text: arXiv Link

References:


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.