

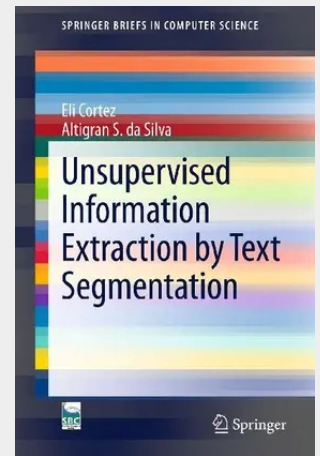
Unsupervised Information Extraction by Text Segmentation

A new unsupervised approach to the problem of Information Extraction by Text Segmentation (IETS) is proposed, implemented and evaluated herein. The authors' approach relies on information available on pre-existing data to learn how to associate segments in the input string with attributes of a given domain relying on a very effective set of content-based features. The effectiveness of the content-based features is also exploited to directly learn from test data structure-based features, with no previous human-driven training, a feature unique to the presented approach. Based on the approach, a number of results are produced to address the IETS problem in an unsupervised fashion. In particular, the authors develop, implement and evaluate distinct IETS methods, namely *ONDUX*, *JUDIE* and *iForm*.

ONDUX (On Demand Unsupervised Information Extraction) is an unsupervised probabilistic approach for IETS that relies on content-based features to bootstrap the learning of structure-based features. *JUDIE* (Joint Unsupervised Structure Discovery and Information Extraction) aims at automatically extracting several semi-structured data records in the form of continuous text and having no explicit delimiters between them. In comparison with other IETS methods, including *ONDUX*, *JUDIE* faces a task considerably harder that is, extracting information while simultaneously uncovering the underlying structure of the implicit records containing it. *iForm* applies the authors' approach to the task of Web form filling. It aims at extracting segments from a data-rich text given as input and associating these segments with fields from a target Web form.

All of these methods were evaluated considering different experimental datasets, which are used to perform a large set of experiments in order to validate the presented approach and methods. These experiments indicate that the proposed approach yields high quality results when compared to state-of-the-art approaches and that it is able to properly support IETS methods in a number of real applications. The findings will prove valuable to practitioners in helping them to understand the current state-of-the-art in unsupervised information extraction techniques, as well as to graduate and undergraduate students of web data management.

A new unsupervised approach to the problem of Information Extraction by Text Segmentation (IETS) is proposed, implemented and evaluated herein. The authors' approach relies on information available on pre-existing data to learn how to associate segments in the input string with attributes of a given domain relying on a very effective set of content-based features. The effectiveness of the content-based features is also exploited to directly learn from test data structure-based features, with no previous human-driven training, a feature unique to the presented approach. Based on the approach, a number of results are produced to address the IETS problem in an unsupervised fashion. In particular, the authors develop, implement and evaluate distinct IETS methods, namely *ONDUX*, *JUDIE* and *iForm*. *ONDUX* (On Demand Unsupervised Information Extraction) is an unsupervised probabilistic approach for IETS that relies on content-based features to bootstrap the learning of structure-based features. *JUDIE* (Joint Unsupervised Structure Discovery and Information Extraction) aims at automatically extracting several semi-structured data records in the form of continuous text and having no explicit delimiters between them. In comparison with other IETS methods, including *ONDUX*, *JUDIE* faces a task considerably harder that is, extracting information while simultaneously uncovering the underlying structure of the implicit records containing it. *iForm* applies the authors' approach to the task of Web form filling. It aims at extracting segments from a data-rich text given as input and associating these segments with fields from a target Web form. All of these methods were evaluated considering different experimental datasets, which are used to perform a large set of experiments in order to validate the presented approach and methods. These experiments indicate that the proposed approach yields high quality results when compared to state-of-the-art approaches and that it is able to properly support IETS methods in a number of real



53,49 €

49,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783319025964

Medium: Buch

ISBN: 978-3-319-02596-4

Verlag: Springer

Erscheinungstermin: 11.11.2013

Sprache(n): Englisch

Auflage: Erscheinungsjahr 2013

Serie: SpringerBriefs in Computer Science

Produktform: Kartoniert

Gewicht: 1825 g

Seiten: 94

Format (B x H): 155 x 235 mm

