Preface

Information Extraction (IE) refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from noisy unstructured textual sources. It derives from the necessity of having unstructured data stored in structured formats (tables, XML), so that it can be further queried, processed, and analyzed.

The IE problem encompasses many distinct sub-problems such as Named Entity Recognition (NER), Open Information Extraction, Relationship Extraction, and Text Segmentation. Information Extraction by Text Segmentation (IETS) is the problem of segmenting unstructured textual inputs to extract implicit data values contained in them.

In this book, we present a novel unsupervised approach for the problem of IETS. This 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 our approach.

Based on this approach, a number of results were obtained to address the IETS problem in an unsupervised fashion. In particular, distinct IETS methods, namely *ONDUX*, *JUDIE*, and *iForm* were implemented, evaluated, and developed.

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. Structure-based features are exploited to disambiguate the extraction of certain attributes through a reinforcement step, which relies on sequencing and positioning of attribute values directly learned *on-demand* from the input texts.

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. In spite of that, it achieves results comparable to the state-of-the-art methods.

iForm applies our 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. The extraction process relies on content-based features learned from data that was previously submitted to the Web form.

All of these methods were evaluated considering different experimental datasets, which we use 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.

Organization

This book consists of seven chapters. Chapter 1 provides an introduction to information extraction (including how information extraction fits into the broader topics of data management), as well as a short description of the main contributions of this book. Chapter 2 then gives an overview of the existing literature and discusses related work.

The core of the book is made up by Chaps. 3–6. Chapter 3 presents the basic concepts and describes an unsupervised approach to exploit pre-existing datasets to support IETS methods. Chapter 4 presents the method called *ONDUX* and all experiments performed to evaluate its performance in comparison to other information extraction methods.

Chapter 5 presents *JUDIE*, an IE method that is able to extract information from text and capable of detecting the structure of each individual records being extracted without any user assistance. Chapter 6 presents *iForm*, a method for dealing with the Web form filling problem that relies on the presented unsupervised proposed. Finally, Chap. 7 presents the conclusions and discusses the future work.

Intended Audience

The aim of this book is to be accessible to researchers, graduate and research students, and to practitioners who work with IE and related areas. It is assumed the reader has some expertise in algorithms and data structures, and web technologies.

This book provides the reader with a broad range of IE concepts and techniques, specifically touching on all aspects of the unsupervised approach presented here. Thus, this book can help researches not only from the IE area, but also from related fields (such as databases, information retrieval, data mining, artificial intelligence, machine learning), as well as students who are interested to enter this field of

research, to become familiar with the recent research developments and identify open research challenges.

This book can help practitioners to better understand the current state-of-the-art in unsupervised IE techniques. Given that in many applications domains it is not feasible to simply use or implement an existing off-the-shelf IE system without substantial adaption, it is crucial for practitioners to understand different aspects of the existing extraction methods. The technical level of this book also makes it accessible to students taking advanced undergraduate and graduate level courses on Web data management.

Acknowledgments

The Authors would like to express their gratitude to colleagues at the UFAM BDRI Group and at the InWeb Project, especially to those we collaborate in many parts of this work: Edleno Moura, Guilherme Toda, Felipe Mesquita, Alberto H. F. Laender, and Marcos Andr′e Gonçalves. We are also grateful to our families and friends for their continuous support.

This work was partially founded by projects DOMAR (CNPq 476798/2011-6), TTDSW (PRONEM/FAPEAM/CNPq), INWeb (MCT/CNPq 57.3871/2008-6), by UOL Bolsa Pesquisa program (grant 20090213165000), and by the authors' individual grants and scholarships from CNPq, CAPES and SUFRAMA.

Manaus, August 2013

Eli Cortez
Altigran S. da Silva

http://www.springer.com/978-3-319-02596-4

Unsupervised Information Extraction by Text Segmentation

Cortez, E.; da Silva, A.

2013, XV, 94 p. 25 illus., Softcover
ISBN: 978-3-319-02596-4