Mining Structures of Factual Knowledge from Text

An Effort-Light Approach

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Mining Human Mobility in Location-Based Social Networks
Huiji Gao and Huan Liu
2015

Mining Latent Entity Structures
Chi Wang and Jiawei Han
2015

Probabilistic Approaches to Recommendations
Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco
2014

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Mining Structures of Factual Knowledge from Text: An Effort-Light Approach
Xiang Ren and Jiawei Han

ISBN: 978-3-031-00784-2 paperback
ISBN: 978-3-031-01912-8 ebook
ISBN: 978-3-031-00107-9 hardcover

DOI 10.1007/978-3-031-01912-8

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #15
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of California, Santa Cruz*
Wei Wang, *University of California, Los Angeles*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Mining Structures of Factual Knowledge from Text

An Effort-Light Approach

Xiang Ren
University of Southern California

Jiawei Han
University of Illinois at Urbana-Champaign

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #15

ABSTRACT

The real-world data, though massive, is largely unstructured, in the form of natural-language text. It is challenging but highly desirable to mine structures from massive text data, without extensive human annotation and labeling. In this book, we investigate the principles and methodologies of mining structures of factual knowledge (e.g., entities and their relationships) from massive, unstructured text corpora.

Departing from many existing structure extraction methods that have heavy reliance on human annotated data for model training, our effort-light approach leverages human-curated facts stored in external knowledge bases as distant supervision and exploits rich data redundancy in large text corpora for context understanding. This effort-light mining approach leads to a series of new principles and powerful methodologies for structuring text corpora, including: (1) entity recognition, typing, and synonym discovery; (2) entity relation extraction; and (3) open-domain attribute-value mining and information extraction. This book introduces this new research frontier and points out some promising research directions.

KEYWORDS

mining factual structures, information extraction, knowledge bases, entity recognition and typing, relation extraction, entity synonym mining, distant supervision, effort-light approach, classification, clustering, real-world applications, scalable algorithms

To my wonderful parents for their love and support.

– Xiang Ren

To my wife Dora, son Lawrence, and grandson Emmett for their love.

– Jiawei Han

Contents

Acknowledgments

The authors would like to acknowledge Wenqi He, Liyuan Liu, Meng Qu, Ellen Wu, Qi Zhu, Jingbo Shang, and Meng Jiang for their tremendous research collaborations.

Han's work was supported in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

Ren's work was sponsored by Google PhD Fellowship, ACM SIGKDD Scholarship, and Richard T. Cheng Endowed Fellowship.

Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any funding agencies.

Xiang Ren and Jiawei Han
June 2018