

# **Mining Structures of Factual Knowledge from Text**

**An Effort-Light Approach**

# Synthesis Lectures on Data Mining and Knowledge Discovery

## 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*

**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 Structures of Factual Knowledge from Text: An Effort-Light Approach

Xiang Ren and Jiawei Han

2018

## Individual and Collective Graph Mining: Principles, Algorithms, and Applications

Danai Koutra and Christos Faloutsos

2017

## Phrase Mining from Massive Text and Its Applications

Jialu Liu, Jingbo Shang, and Jiawei Han

2017

## Exploratory Causal Analysis with Time Series Data

James M. McCracken

2016

## 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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2018

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

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

	<b>Acknowledgments</b> .....	<b>xv</b>
<b>1</b>	<b>Introduction</b> .....	<b>1</b>
1.1	Overview of the Book .....	2
1.1.1	Part I: Identifying Typed Entities .....	5
1.1.2	Part II: Extracting Typed Entity Relationships .....	10
1.1.3	Part III: Toward Automated Factual Structure Mining .....	14
<b>2</b>	<b>Background</b> .....	<b>19</b>
2.1	Entity Structures .....	19
2.2	Relation Structures .....	21
2.3	Distant Supervision from Knowledge Bases .....	21
2.4	Mining Entity and Relation Structures .....	23
2.5	Common Notations .....	24
<b>3</b>	<b>Literature Review</b> .....	<b>25</b>
3.1	Hand-Crafted Methods .....	25
3.2	Traditional Supervised Learning Methods .....	26
3.2.1	Sequence Labeling Methods .....	27
3.2.2	Supervised Relation Extraction Methods .....	28
3.3	Weakly Supervised Extraction Methods .....	28
3.3.1	Semi-Supervised Learning .....	29
3.3.2	Pattern-Based Bootstrapping .....	29
3.4	Distantly Supervised Learning Methods .....	30
3.5	Learning with Noisy Labeled Data .....	30
3.6	Open-Domain Information Extraction .....	31



	<b>PART I Identifying Typed Entities . . . . .</b>	<b>33</b>
<b>4</b>	<b>Entity Recognition and Typing with Knowledge Bases . . . . .</b>	<b>35</b>
4.1	Overview and Motivation . . . . .	35
4.2	Problem Definition . . . . .	38
4.3	Relation Phrase-Based Graph Construction . . . . .	40
4.3.1	Candidate Generation . . . . .	40
4.3.2	Mention-Name Subgraph . . . . .	41
4.3.3	Name-Relation Phrase Subgraph . . . . .	42
4.3.4	Mention Correlation Subgraph . . . . .	43
4.4	Clustering-Integrated Type Propagation on Graphs . . . . .	44
4.4.1	Seed Mention Generation . . . . .	45
4.4.2	Relation Phrase Clustering . . . . .	45
4.4.3	The Joint Optimization Problem . . . . .	46
4.4.4	The ClusType Algorithm . . . . .	48
4.4.5	Computational Complexity Analysis . . . . .	50
4.5	Experiments . . . . .	51
4.5.1	Data Preparation . . . . .	51
4.5.2	Experimental Settings . . . . .	52
4.5.3	Experiments and Performance Study . . . . .	53
4.6	Discussion . . . . .	56
4.7	Summary . . . . .	58
<b>5</b>	<b>Fine-Grained Entity Typing with Knowledge Bases . . . . .</b>	<b>59</b>
5.1	Overview and Motivation . . . . .	59
5.2	Preliminaries . . . . .	62
5.3	The AFET Framework . . . . .	64
5.3.1	Text Feature Generation . . . . .	64
5.3.2	Training Set Partition . . . . .	64
5.3.3	The Joint Mention-Type Model . . . . .	65
5.3.4	Modeling Type Correlation . . . . .	65
5.3.5	Modeling Noisy Type Labels . . . . .	67
5.3.6	Hierarchical Partial-Label Embedding . . . . .	68
5.4	Experiments . . . . .	69
5.4.1	Data Preparation . . . . .	69
5.4.2	Evaluation Settings . . . . .	70
5.4.3	Performance Comparison and Analyses . . . . .	70

5.5	Discussion and Case Analysis	71
5.6	Summary	72
<b>6</b>	<b>Synonym Discovery from Large Corpus</b>	<b>75</b>
	<i>Meng Qu</i>	
	<i>Department of Computer Science, University of Illinois at Urbana-Champaign</i>	
6.1	Overview and Motivation	75
6.1.1	Challenges	76
6.1.2	Proposed Solution	78
6.2	The DPE Framework	78
6.2.1	Synonym Seed Collection	79
6.2.2	Joint Optimization Problem	79
6.2.3	Distributional Module	80
6.2.4	Pattern Module	80
6.3	Experiment	81
6.4	Summary	84
	 <b>PART II Extracting Typed Relationships</b>	 <b>85</b>
<b>7</b>	<b>Joint Extraction of Typed Entities and Relationships</b>	<b>87</b>
7.1	Overview and Motivation	87
7.2	Preliminaries	90
7.3	The CoType Framework	92
7.3.1	Candidate Generation	93
7.3.2	Joint Entity and Relation Embedding	95
7.3.3	Model Learning and Type Inference	101
7.4	Experiments	102
7.4.1	Data Preparation and Experiment Setting	102
7.4.2	Experiments and Performance Study	105
7.5	Discussion	107
7.6	Summary	110
<b>8</b>	<b>Pattern-Enhanced Embedding Learning for Relation Extraction</b>	<b>111</b>
	<i>Meng Qu</i>	
	<i>Department of Computer Science, University of Illinois at Urbana-Champaign</i>	
8.1	Overview and Motivation	112

8.1.1	Challenges	112
8.1.2	Proposed Solution	113
8.2	The REPEL Framework	114
8.3	Experiment	115
8.4	Summary	118
<b>9</b>	<b>Heterogeneous Supervision for Relation Extraction</b>	<b>119</b>
	<i>Liyuan Liu</i>	
	<i>Department of Computer Science, University of Illinois at Urbana-Champaign</i>	
9.1	Overview and Motivation	119
9.2	Preliminaries	120
9.2.1	Relation Extraction	120
9.2.2	Heterogeneous Supervision	121
9.2.3	Problem Definition	121
9.3	The REHESSION Framework	121
9.3.1	Modeling Relation Mention	122
9.3.2	True Label Discovery	124
9.3.3	Modeling Relation Type	125
9.3.4	Model Learning	125
9.3.5	Relation Type Inference	126
9.4	Experiments	126
9.5	Summary	127
<b>10</b>	<b>Indirect Supervision: Leveraging Knowledge from Auxiliary Tasks</b>	<b>129</b>
	<i>Zegiu Wu</i>	
	<i>Department of Computer Science, University of Illinois at Urbana-Champaign</i>	
10.1	Overview and Motivation	129
10.1.1	Challenges	130
10.1.2	Proposed Solution	131
10.2	The Proposed Approach	132
10.2.1	Heterogeneous Network Construction	132
10.2.2	Joint RE and QA Embedding	133
10.2.3	Type Inference	134
10.3	Experiments	134
10.4	Summary	136

## PART III Toward Automated Factual Structure Mining . . . . . 137

<b>11</b>	<b>Mining Entity Attribute Values with Meta Patterns . . . . .</b>	<b>139</b>
	<i>Meng Jiang</i>	
	<i>Department of Computer Science and Engineering, University of Notre Dame</i>	
11.1	Overview and Motivation . . . . .	139
11.1.1	Challenges . . . . .	140
11.1.2	Proposed Solution . . . . .	140
11.1.3	Problem Formulation . . . . .	141
11.2	The MetaPAD Framework . . . . .	142
11.2.1	Generating Meta Patterns by Context-Aware Segmentation . . . . .	142
11.2.2	Grouping Synonymous Meta Patterns . . . . .	144
11.2.3	Adjusting Type Levels for Preciseness . . . . .	145
11.3	Summary . . . . .	145
<b>12</b>	<b>Open Information Extraction with Global Structure Cohesiveness . . . . .</b>	<b>147</b>
	<i>Qi Zhu</i>	
	<i>Department of Computer Science, University of Illinois at Urbana-Champaign</i>	
12.1	Overview and Motivation . . . . .	147
12.1.1	Proposed Solution . . . . .	149
12.2	The ReMine Framework . . . . .	150
12.2.1	The Joint Optimization Problem . . . . .	151
12.3	Summary . . . . .	151
<b>13</b>	<b>Applications . . . . .</b>	<b>153</b>
13.1	Structuring Life Science Papers: The Life-iNet System . . . . .	153
13.2	Extracting Document Facets from Technical Corpora . . . . .	156
13.3	Comparative Document Analysis . . . . .	157
<b>14</b>	<b>Conclusions . . . . .</b>	<b>161</b>
14.1	Effort-Light StructMine: Summary . . . . .	161
14.2	Conclusion . . . . .	163
<b>15</b>	<b>Vision and Future Work . . . . .</b>	<b>165</b>
15.1	Extracting Implicit Patterns from Massive Unlabeled Corpora . . . . .	165
15.2	Enriching Factual Structure Representation . . . . .	165

**Bibliography .....167**

**Authors' Biographies .....183**

# 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](http://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