

# **Phrase Mining from Massive Text and Its Applications**

# 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.

## 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 2017

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.

Phrase Mining from Massive Text and Its Applications

Jialu Liu, Jingbo Shang, and Jiawei Han

ISBN: 978-3-031-00782-8      paperback

ISBN: 978-3-031-01910-4      ebook

DOI 10.1007/978-3-031-01910-4

A Publication in the Springer series

*SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY*

Lecture #13

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

# Phrase Mining from Massive Text and Its Applications

Jialu Liu  
Google

Jingbo Shang  
University of Illinois at Urbana-Champaign

Jiawei Han  
University of Illinois at Urbana-Champaign

*SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE  
DISCOVERY #13*

## ABSTRACT

A lot of digital ink has been spilled on “big data” over the past few years. Most of this surge owes its origin to the various types of unstructured data in the wild, among which the proliferation of text-heavy data is particularly overwhelming, attributed to the daily use of web documents, business reviews, news, social posts, etc., by so many people worldwide. A core challenge presents itself: How can one efficiently and effectively turn massive, unstructured text into structured representation so as to further lay the foundation for many other downstream text mining applications?

In this book, we investigated one promising paradigm for representing unstructured text, that is, through automatically identifying high-quality phrases from innumerable documents. In contrast to a list of frequent  $n$ -grams without proper filtering, users are often more interested in results based on variable-length phrases with certain semantics such as scientific concepts, organizations, slogans, and so on. We propose new principles and powerful methodologies to achieve this goal, from the scenario where a user can provide meaningful guidance to a fully automated setting through distant learning. This book also introduces applications enabled by the mined phrases and points out some promising research directions.

## KEYWORDS

phrase mining, phrase quality, phrasal segmentation, distant supervision, text mining, real-world applications, efficient and scalable algorithms

# Contents

|          |   |           |
|----------|---|-----------|
|          | <b>Acknowledgments</b> .....                          | <b>ix</b> |
| <b>1</b> | <b>Introduction</b> .....                             | <b>1</b>  |
| 1.1      | Motivation .....                                      | 1         |
| 1.2      | What is Phrase Mining? .....                          | 2         |
| 1.3      | Outline of the Book .....                             | 3         |
| <b>2</b> | <b>Quality Phrase Mining with User Guidance</b> ..... | <b>5</b>  |
| 2.1      | Overview .....  | 5         |
| 2.2      | Phrasal Segmentation .....                            | 7         |
| 2.3      | Supervised Phrase Mining Framework .....              | 9         |
| 2.3.1    | Frequent Phrase Detection .....                       | 10        |
| 2.3.2    | Phrase Quality Estimation .....                       | 10        |
| 2.3.3    | Rectification through Phrasal Segmentation .....      | 14        |
| 2.3.4    | Feedback as Segmentation Features .....               | 17        |
| 2.3.5    | Complexity Analysis .....                             | 19        |
| 2.4      | Experimental Study .....                              | 20        |
| 2.4.1    | Quantitative Evaluation and Results .....             | 21        |
| 2.4.2    | Model Selection .....                                 | 24        |
| 2.4.3    | Efficiency Study .....                                | 27        |
| 2.4.4    | Case Study .....                                      | 28        |
| 2.5      | Summary .....   | 30        |
| <b>3</b> | <b>Automated Quality Phrase Mining</b> .....          | <b>35</b> |
| 3.1      | Overview .....  | 35        |
| 3.2      | Automated Phrase Mining Framework .....               | 37        |
| 3.2.1    | Phrase Label Generation .....                         | 38        |
| 3.2.2    | Phrase Quality Estimation .....                       | 40        |
| 3.2.3    | POS-guided Phrasal Segmentation .....                 | 41        |
| 3.2.4    | Phrase Quality Re-estimation .....                    | 45        |
| 3.2.5    | Complexity Analysis .....                             | 46        |
| 3.3      | Experimental Study .....                              | 46        |

|          |   |           |
|----------|---|-----------|
| 3.3.1    | Experimental Settings . . . . .                     | 48        |
| 3.3.2    | Quantitative Evaluation and Results . . . . .       | 48        |
| 3.3.3    | Distant Training Exploration . . . . .              | 50        |
| 3.3.4    | POS-guided Phrasal Segmentation . . . . .           | 52        |
| 3.3.5    | Efficiency Study . . . . .                          | 53        |
| 3.3.6    | Case Study . . . . .                                | 53        |
| <b>4</b> | <b>Phrase Mining Applications . . . . .</b>         | <b>55</b> |
| 4.1      | Latent Keyphrase Inference . . . . .                | 55        |
| 4.2      | Topic Exploration for Document Collection . . . . . | 60        |
| 4.3      | Knowledge Base Construction . . . . .               | 68        |
| 4.4      | Research Frontier . . . . .                         | 71        |
|          | <b>Bibliography . . . . .</b>                       | <b>73</b> |
|          | <b>Authors' Biographies . . . . .</b>               | <b>79</b> |



# Acknowledgments

The authors would like to acknowledge Xiang Ren, Fangbo Tao, and Huan Gui for their contribution to Chapter 4.

The research was supported in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617, and IIS 16-18481, 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)). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. The views and conclusions contained in our research publications are those of the authors and should not be interpreted as representing any funding agencies.

Jialu Liu, Jingbo Shang, and Jiawei Han  
February 2017