Phrase Mining from Massive Text and Its Applications

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.

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

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

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). 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