Synthesis Lectures on
Data Mining and Knowledge Discovery

SYNTHESIS
COLLECTION OF TECHNOLOGY

Jiaming Shen · Jiawei Han

Automated Taxonomy Discovery and Exploration

Springer

Synthesis Lectures on Data Mining and Knowledge Discovery

The series focuses 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.

Jiaming Shen · Jiawei Han

Automated Taxonomy Discovery and Exploration

Springer

Jiaming Shen
Google Research
New York City, NY, USA

Jiawei Han
University of Illinois Urbana-Champaign
Urbana, IL, USA

ISSN 2151-0067 ISSN 2151-0075 (electronic)
Synthesis Lectures on Data Mining and Knowledge Discovery
ISBN 978-3-031-11404-5 ISBN 978-3-031-11405-2 (eBook)
https://doi.org/10.1007/978-3-031-11405-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my family for their love and support.

Preface

In today's information era, people are inundated with vast amounts of text data. Every day, there are thousands of scientific papers, tens of thousands of news articles, corporate reports, and millions of social media posts produced and shared worldwide. Turning those massive text data into actionable knowledge is an essential research issue in data science and lays the foundation for realizing machine intelligence.

In this book, we discuss how to unleash hidden knowledge buried in unstructured text. We propose to first structure raw text using taxonomies and then analyze structured text in a more fine-grained and semantic way. Due to the diversity of application scenarios, different corpora or different use cases may call for different taxonomies. For example, one analyst aiming to find experts in different scientific areas may want a field-of-study taxonomy, while another analyst who studies the technology readiness may call for a taxonomy capturing technology dependencies. Moreover, even within one taxonomy, we also enable users to organize concepts at their will, such as with different levels containing concepts of different categories. For instance, in a computer science taxonomy, top levels could be about *field of studies*, intermediate levels may discuss *research tasks*, and the bottom levels can cover *evaluation metrics*. Asking human experts to manually curate those taxonomies, one for every possible application, is time-consuming, costly, and unscalable. Therefore, we propose to automatically discover and explore taxonomies based on the datasets and applications, with critical but minimal human guidance.

This book outlines a data-driven approach that automatically constructs, enriches, and applies taxonomies for unleashing knowledge from massive unstructured text. Particularly, we investigate four areas of research, including:

1. **Concept Set Discovery.** To obtain concept nodes in the taxonomy, we first develop a collection of concept set expansion methods to extract concepts from text corpora by expanding a small set of seed concepts into a complete list of concepts that belong to the same semantic class.
2. **Taxonomy Construction.** To organize above identified concepts into hierarchical structure, we propose a set of taxonomy construction methods to discover taxonomic relations among concepts by analyzing example relation instances (i.e. , concept pairs

indicating the target relation semantics) and utilizing distant supervision from existing, open-domain knowledge bases.
3. **Taxonomy Enrichment.** As human knowledge is constantly growing, a static taxonomy may fail to capture emerging user needs. Thus, a taxonomy enrichment step would be essential to keep our taxonomies up-to-date in real-world applications. We facilitate this process by expanding the taxonomy to incorporate new concepts.
4. **Taxonomy-Guided Classification.** After an up-to-date taxonomy is obtained, we develop principled methods to leverage taxonomies for classification tasks.

Together, these pieces constitute an integrated framework for leveraging taxonomies to convert massive text data into actionable knowledge.

New York City, USA Jiaming Shen

Contents