Synthesis Lectures on Data Mining and Knowledge Discovery

Series Editors

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

Lise Getoor, University of California, Santa Cruz, Santa Cruz, USA

Johannes Gehrke, Microsoft Corporation, Redmond, WA, USA

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

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

1 Introduction 1
1.1 Overview 1
1.2 Technical Roadmap 4
1.2.1 Concept Set Expansion 4
1.2.2 Taxonomy Construction 4
1.2.3 Taxonomy Enrichment 5
1.2.4 Taxonomy-Guided Classification 5
1.3 Organization 6
References 6

2 Concept Set Expansion 9
2.1 Overview and Motivations 9
2.2 Related Work 11
2.3 SetExpan: Weakly-Supervised Concept Set Expansion 12
2.3.1 Data Model and Context Features 12
2.3.2 Context-Dependent Concept Similarity 14
2.3.3 Context Feature Selection 14
2.3.4 Concept Selection via Rank Ensemble 15
2.4 Experiments 17
2.4.1 Datasets 17
2.4.2 Compared Methods 18
2.4.3 Evaluation Metrics 19
2.4.4 Overall Performance 19
2.4.5 Ablation Studies 21
2.4.6 Case Studies 23
2.5 Extensions of SetExpan 24
2.5.1 Addressing Concept Drifts via Auxiliary Sets Generation and Co-expansion 24
2.5.2 Probing Knowledge from Pre-trained Language Models 25

2.6 Summary 27
References 27

3 Taxonomy Construction 31
3.1 Overview and Motivations 31
3.2 Related Work 33
3.3 HiExpan: Task-Guided Concept Taxonomy Construction 34
3.3.1 Problem Formulation 34
3.3.2 Framework Overview 35
3.3.3 Key Term Extraction 35
3.3.4 Iterative Width and Depth Expansion 36
3.3.5 Taxonomy Global Optimization 39
3.4 Experiments 40
3.4.1 Datasets 40
3.4.2 Compared Methods 41
3.4.3 Evaluation Metrics 41
3.4.4 Quantitative Results 42
3.4.5 Case Studies 43
3.5 Summary 44
References 45

4 Taxonomy Enrichment 49
4.1 Overview and Motivations 49
4.2 Related Work 51
4.3 TaxoExpan: Self-supervised Taxonomy Expansion 52
4.3.1 Problem Formulation 52
4.3.2 Taxonomy Modeling and Expansion Goal 53
4.3.3 Query-Anchor Matching Model 54
4.3.4 Model Learning and Inference 58
4.4 Experiments 61
4.4.1 Experiments on MAG Dataset 61
4.4.2 Experiments on SemEval Dataset 69
4.5 Extensions of TaxoExpan 71
4.5.1 Incorporating More Fine-Grained Self-supervision Tasks 71
4.5.2 Identifying Potential Children Concepts 73
4.5.3 Modeling Relations Among News Concepts 75
4.6 Summary 78
References 79

5 Taxonomy-Guided Classification 83
5.1 Overview and Motivations 83
5.2 Related Work 85

5.3 TaxoClass: Weakly-Supervised Hierarchical Multi-label Text Classification . . . 86
5.3.1 Problem Formulation . . . 86
5.3.2 Document-Class Similarity Calculation . . . 87
5.3.3 Document Core Class Mining . . . 87
5.3.4 Core Class Guided Classifier Training . . . 89
5.3.5 Multi-label Self-training . . . 91
5.4 Experiments . . . 92
5.4.1 Datasets . . . 92
5.4.2 Compared Methods . . . 92
5.4.3 Evaluation Metrics . . . 94
5.4.4 Implementation Details . . . 94
5.4.5 Overall Performance Comparison . . . 95
5.4.6 Effectiveness of Core Class Mining . . . 95
5.4.7 Analysis of Classifier Architecture . . . 96
5.4.8 Supervision Signals in Class Names . . . 96
5.5 Summary . . . 97
References . . . 98

6 Conclusions . . . 101
6.1 Summary . . . 101
6.2 Future Work . . . 101
6.2.1 Integrate Heterogeneous Modalities and Sources . . . 102
6.2.2 Engage with Human Behaviors and Interactions . . . 102
6.2.3 Preserve Data Privacy and Model Security . . . 103
References . . . 103