Multidimensional Mining of Massive Text Data

Synthesis Lectures on Data Mining and Knowledge Discovery

Editors
Jiawei Han, ***University of Illinois at Urbana-Champaign***
Johannes Gehrke, ***Cornell University***
Lise Getoor, ***University of California, Santa Cruz***
Robert Grossman, ***University of Chicago***
Wei Wang, ***University of California, Los Angeles***

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.

Multidimensional Mining of Massive Text Data
Chao Zhang and Jiawei Han
2019

Exploiting the Power of Group Differences: Using Patterns to Solve Data Analysis Problems
Guozhu Dong
2018

Mining Structures of Factual Knowledge from Text
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 2019

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.

Multidimensional Mining of Massive Text Data

Chao Zhang and Jiawei Han

ISBN: 978-3-031-00786-6 paperback
ISBN: 978-3-031-01914-2 ebook
ISBN: 978-3-031-00109-3 hardcover

DOI 10.1007/978-3-031-01914-2

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #17

Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Johannes Gehrke, *Cornell University*
Lise Getoor, *University of California, Santa Cruz*
Robert Grossman, *University of Chicago*
Wei Wang, *University of California, Los Angeles*

Series ISSN
Print 2151-0067 Electronic 2151-0075

Multidimensional Mining of Massive Text Data

Chao Zhang
Georgia Institute of Technology

Jiawei Han
University of Illinois at Urbana-Champaign

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #17

ABSTRACT

Unstructured text, as one of the most important data forms, plays a crucial role in data-driven decision making in domains ranging from social networking and information retrieval to scientific research and healthcare informatics. In many emerging applications, people's information need from text data is becoming multidimensional—they demand useful insights along multiple aspects from a text corpus. However, acquiring such multidimensional knowledge from massive text data remains a challenging task.

This book presents data mining techniques that turn unstructured text data into multidimensional knowledge. We investigate two core questions. (1) How does one identify task-relevant text data with declarative queries in multiple dimensions? (2) How does one distill knowledge from text data in a multidimensional space? To address the above questions, we develop a text cube framework. First, we develop a cube construction module that organizes unstructured data into a cube structure, by discovering latent multidimensional and multi-granular structure from the unstructured text corpus and allocating documents into the structure. Second, we develop a cube exploitation module that models multiple dimensions in the cube space, thereby distilling from user-selected data multidimensional knowledge. Together, these two modules constitute an integrated pipeline: leveraging the cube structure, users can perform multidimensional, multigranular data selection with declarative queries; and with cube exploitation algorithms, users can extract multidimensional patterns from the selected data for decision making.

The proposed framework has two distinctive advantages when turning text data into multidimensional knowledge: flexibility and label-efficiency. First, it enables acquiring multidimensional knowledge flexibly, as the cube structure allows users to easily identify task-relevant data along multiple dimensions at varied granularities and further distill multidimensional knowledge. Second, the algorithms for cube construction and exploitation require little supervision; this makes the framework appealing for many applications where labeled data are expensive to obtain.

KEYWORDS

text mining, multidimensional analysis, data cube, limited supervision

Contents

1 Introduction . . . 1
1.1 Overview . . . 1
1.2 Main Parts . . . 3
1.2.1 Part I: Cube Construction . . . 3
1.2.2 Part II: Cube Exploitation . . . 5
1.2.3 Example Applications . . . 5
1.3 Technical Roadmap . . . 6
1.3.1 Task 1: Taxonomy Generation . . . 7
1.3.2 Task 2: Document Allocation . . . 8
1.3.3 Task 3: Multidimensional Summarization . . . 9
1.3.4 Task 4: Cross-Dimension Prediction . . . 9
1.3.5 Task 5: Abnormal Event Detection . . . 9
1.3.6 Summary . . . 10
1.4 Organization . . . 10

PART I Cube Construction Algorithms . . . 11

2 Topic-Level Taxonomy Generation . . . 13
2.1 Overview . . . 13
2.2 Related Work . . . 16
2.2.1 Supervised Taxonomy Learning . . . 16
2.2.2 Pattern-Based Extraction . . . 16
2.2.3 Clustering-Based Taxonomy Construction . . . 17
2.3 Preliminaries . . . 18
2.3.1 Problem Definition . . . 18
2.3.2 Method Overview . . . 18
2.4 Adaptive Term Clustering . . . 18
2.4.1 Spherical Clustering for Topic Splitting . . . 19
2.4.2 Identifying Representative Terms . . . 20
2.5 Adaptive Term Embedding . . . 22

2.5.1 Distributed Term Representations 22
2.5.2 Learning Local Term Embeddings 22
2.6 Experimental Evaluation 23
2.6.1 Experimental Setup 23
2.6.2 Qualitative Results 25
2.6.3 Quantitative Analysis 27
2.7 Summary 30

3 Term-Level Taxonomy Generation 31
Jiaming Shen
University of Illinois at Urbana-Champaign
3.1 Overview 31
3.2 Related Work 33
3.3 Problem Formulation 34
3.4 The HiExpan Framework 34
3.4.1 Framework Overview 34
3.4.2 Key Term Extraction 35
3.4.3 Hierarchical Tree Expansion 35
3.4.4 Taxonomy Global Optimization 40
3.5 Experiments 42
3.5.1 Experimental Setup 42
3.5.2 Qualitative Results 43
3.5.3 Quantitative Results 45
3.6 Summary 48

4 Weakly Supervised Text Classification 49
Yu Meng
University of Illinois at Urbana-Champaign
4.1 Overview 49
4.2 Related Work 51
4.2.1 Latent Variable Models 51
4.2.2 Embedding-Based Models 52
4.3 Preliminaries 52
4.3.1 Problem Formulation 53
4.3.2 Method Overview 53
4.4 Pseudo-Document Generation 53

4.4.1 Modeling Class Distribution 54
4.4.2 Generating Pseudo-Documents 55
4.5 Neural Models with Self-Training 57
4.5.1 Neural Model Pre-training 57
4.5.2 Neural Model Self-Training 57
4.5.3 Instantiating with CNNs and RNNs 58
4.6 Experiments 59
4.6.1 Datasets 60
4.6.2 Baselines 60
4.6.3 Experiment Settings 61
4.6.4 Experiment Results 62
4.6.5 Parameter Study 64
4.6.6 Case Study 68
4.7 Summary 69

5 Weakly Supervised Hierarchical Text Classification 71

Yu Meng

University of Illinois at Urbana-Champaign

5.1 Overview 71
5.2 Related Work 73
5.2.1 Weakly Supervised Text Classification 73
5.2.2 Hierarchical Text Classification 73
5.3 Problem Formulation 74
5.4 Pseudo-Document Generation 74
5.5 The Hierarchical Classification Model 77
5.5.1 Local Classifier Pre-Training 77
5.5.2 Global Classifier Self-Training 77
5.5.3 Blocking Mechanism 79
5.5.4 Inference 79
5.5.5 Algorithm Summary 79
5.6 Experiments 80
5.6.1 Experiment Settings 80
5.6.2 Quantitative Comparision 83
5.6.3 Component-Wise Evaluation 83
5.7 Summary 87

PART II Cube Exploitation Algorithms 89

6 Multidimensional Summarization 91

Fangbo Tao

Facebook Inc.

6.1 Introduction 91
6.2 Related Work 94
6.3 Preliminaries 95
6.3.1 Text Cube Preliminaries 95
6.3.2 Problem Definition 96
6.4 The Ranking Measure 97
6.4.1 Popularity and Integrity 97
6.4.2 Neighborhood-Aware Distinctiveness 98
6.5 The RepPhrase Method 101
6.5.1 Overview 101
6.5.2 Hybrid Offline Materialization 102
6.5.3 Optimized Online Processing 106
6.6 Experiments 107
6.6.1 Experimental Setup 107
6.6.2 Effectiveness Evaluation 108
6.6.3 Efficiency Evaluation 112
6.7 Summary 116

7 Cross-Dimension Prediction in Cube Space 117

7.1 Overview 117
7.2 Related Work 119
7.3 Preliminaries 120
7.3.1 Problem Description 120
7.3.2 Method Overview 120
7.4 Semi-Supervised Multimodal Embedding 122
7.4.1 The Unsupervised Reconstruction Task 122
7.4.2 The Supervised Classification Task 124
7.4.3 The Optimization Procedure 124
7.5 Online Updating of Multimodal Embedding 125
7.5.1 Life-Decaying Learning 125
7.5.2 Constraint-Based Learning 126

7.5.3 Complexity Analysis . . . 129
7.6 Experiments . . . 130
7.6.1 Experimental Setup . . . 130
7.6.2 Quantitative Comparison . . . 133
7.6.3 Case Studies . . . 134
7.6.4 Effects of Parameters . . . 137
7.6.5 Downstream Application . . . 139
7.7 Summary . . . 141

8 Event Detection in Cube Space . . . 143
8.1 Overview . . . 143
8.2 Related Work . . . 145
8.2.1 Bursty Event Detection . . . 145
8.2.2 Spatiotemporal Event Detection . . . 146
8.3 Preliminaries . . . 147
8.3.1 Problem Definition . . . 147
8.3.2 Method Overview . . . 147
8.3.3 Multimodal Embedding . . . 149
8.4 Candidate Generation . . . 150
8.4.1 A Bayesian Mixture Clustering Model . . . 151
8.4.2 Parameter Estimation . . . 152
8.5 Candidate Classification . . . 153
8.5.1 Features Induced from Multimodal Embeddings . . . 153
8.5.2 The Classification Procedure . . . 154
8.6 Supporting Continuous Event Detection . . . 154
8.7 Complexity Analysis . . . 154
8.8 Experiments . . . 155
8.8.1 Experimental Settings . . . 155
8.8.2 Qualitative Results . . . 157
8.8.3 Quantitative Results . . . 160
8.8.4 Scalability Study . . . 161
8.8.5 Feature Importance . . . 162
8.9 Summary . . . 162

9 Conclusions . . . 165
9.1 Summary . . . 165
9.2 Future Work . . . 166

Bibliography 169

Authors' Biographies 183