Studies in Big Data

Volume 45

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and others. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

**Indexing: The books of this series are submitted to ISI Web of Science, DBLP, Ulrichs, MathSciNet, Current Mathematical Publications, Mathematical Reviews, Zentralblatt Math: MetaPress and Springerlink.

More information about this series at http://www.springer.com/series/11970

Taeho Jo

Text Mining

Concepts, Implementation, and Big Data Challenge

Springer

Taeho Jo
School of Game
Hongik University
Seoul, Korea (Republic of)

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-319-91814-3 ISBN 978-3-319-91815-0 (eBook)
https://doi.org/10.1007/978-3-319-91815-0

Library of Congress Control Number: 2018942701

© Springer International Publishing AG, part of Springer Nature 2019
This work is subject to copyright. All rights are reserved 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, express 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.

Printed on acid-free paper

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

Preface

This book is concerned with the concept, the theories, and the implementations of text mining. In Part I, we provide the fundamental knowledge about the text mining tasks, such as text preprocessing and text association. In Parts II and III, we describe the representative approaches to text mining tasks, provide guides for implementing the text mining systems by presenting the source codes in Java, and explain the schemes of evaluating them. In Part IV, we cover other text mining tasks, such as text summarization and segmentation, and the composite text mining tasks. Overall, in this book, we explain the text mining tasks in the functional view, describe the main approaches, and provide guidance for implementing systems, mainly about text categorization and clustering.

There are a few factors that provided motivation for writing this book. Textual data including web documents were already dominant over relational data items in information systems, strongly, since the 1990s. So we need techniques of mining important knowledge from textual data as well as relational data. Once obtaining the techniques, we need to provide guides for implementing text mining systems for system developers. Therefore, in this book, we set the functional views of the text mining tasks, the approaches to them, and the implementation of the text mining systems as the scope of this book.

This book consists of four parts: foundation (Chaps. 1–4), text categorization (Chaps. 5–8), text clustering (Chaps. 9–12), and advanced topics (Chaps. 13–16). In the first part of this book, we provide the introduction to text mining and the schemes of preprocessing texts before performing the text mining tasks. In the second part, we cover types of text categorization, the approaches to the task, and the schemes of implementing and evaluating the text categorization systems. In the third part, we describe text clustering with respect to its types, approaches, implementations, and evaluation schemes. In the last part, we consider other text mining tasks such as text summarization and segmentation, and combinations of the text mining tasks.

This book is intended for the three groups of readers: students, professors, and researchers. Senior undergraduate and graduate students are able to study the contents of text mining by themselves with this book. For lecturers and professors

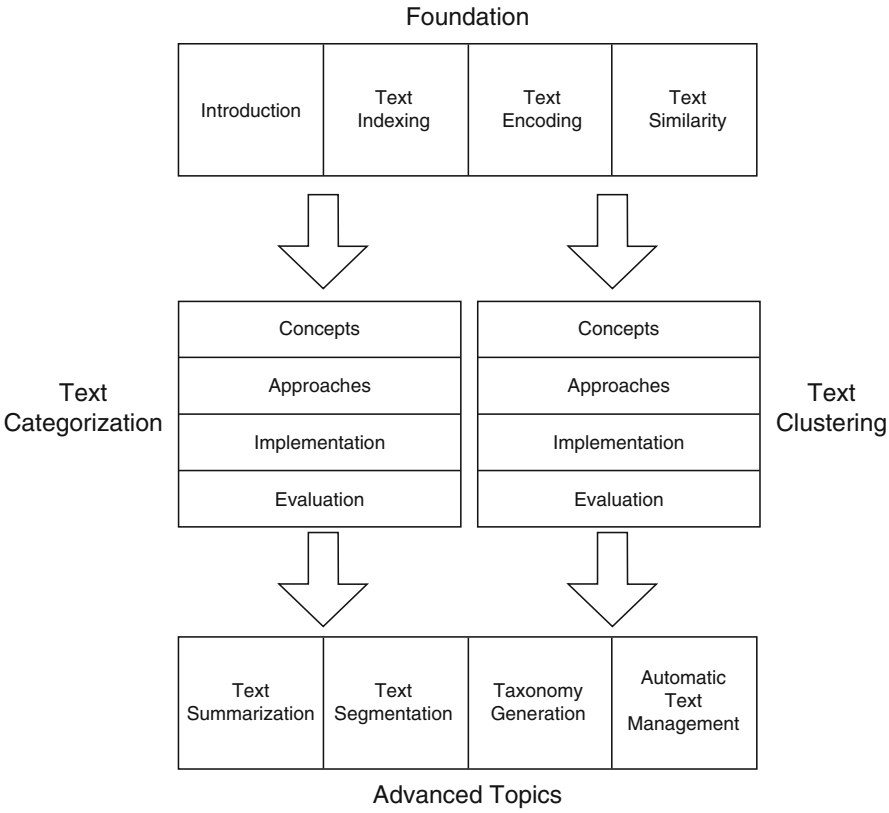

Fig. 1 Organization of four parts

in universities, this book may be used as the main material for providing lectures on text mining. For researchers and system developers in industrial organizations, it may be used as the guide book for developing text mining systems. For understanding this book, it requires only elementary level of the three disciplines of mathematics: probabilities, linear algebra, and vector calculus.

As illustrated in Fig. 1, this book is organized with the four parts. Part I deals with the four aspects: introduction, text indexing, text encoding, and text association. Part II focuses on the text categorization with respect to its concepts, approaches, implementations, and evaluation schemes. Part III mention the text clustering with respect to its four aspects. Part IV is concerned with other text mining tasks and the hybrid ones, as the advanced topics.

Sejong, Korea Taeho Jo

Contents

Part I Foundation

1 Introduction 3
1.1 Definition of Text Mining 3
1.2 Texts 4
1.2.1 Text Components 5
1.2.2 Text Formats 6
1.3 Data Mining Tasks 7
1.3.1 Classification 7
1.3.2 Clustering 9
1.3.3 Association 10
1.4 Data Mining Types 11
1.4.1 Relational Data Mining 12
1.4.2 Web Mining 13
1.4.3 Big Data Mining 14
1.5 Summary 16

2 Text Indexing 19
2.1 Overview of Text Indexing 19
2.2 Steps of Text Indexing 21
2.2.1 Tokenization 21
2.2.2 Stemming 23
2.2.3 Stop-Word Removal 24
2.2.4 Term Weighting 25
2.3 Text Indexing: Implementation 27
2.3.1 Class Definition 27
2.3.2 Stemming Rule 30
2.3.3 Method Implementations 32

2.4 Additional Steps 35
2.4.1 Index Filtering 35
2.4.2 Index Expansion 37
2.4.3 Index Optimization 38
2.5 Summary 40

3 Text Encoding 41
3.1 Overview of Text Encoding 41
3.2 Feature Selection 43
3.2.1 Wrapper Approach 43
3.2.2 Principal Component Analysis 44
3.2.3 Independent Component Analysis 46
3.2.4 Singular Value Decomposition 49
3.3 Feature Value Assignment 50
3.3.1 Assignment Schemes 50
3.3.2 Similarity Computation 52
3.4 Issues of Text Encoding 53
3.4.1 Huge Dimensionality 53
3.4.2 Sparse Distribution 54
3.4.3 Poor Transparency 55
3.5 Summary 57

4 Text Association 59
4.1 Overview of Text Association 59
4.2 Data Association 61
4.2.1 Functional View 61
4.2.2 Support and Confidence 62
4.2.3 Apriori Algorithm 64
4.3 Word Association 66
4.3.1 Word Text Matrix 66
4.3.2 Functional View 68
4.3.3 Simple Example 69
4.4 Text Association 71
4.4.1 Functional View 71
4.4.2 Simple Example 72
4.5 Overall Summary 74

Part II Text Categorization

5 Text Categorization: Conceptual View 79
5.1 Definition of Text Categorization 79
5.2 Data Classification 81
5.2.1 Binary Classification 81
5.2.2 Multiple Classification 82
5.2.3 Classification Decomposition 83
5.2.4 Regression 85

5.3 Classification Types 86
5.3.1 Hard vs Soft Classification 86
5.3.2 Flat vs Hierarchical Classification 88
5.3.3 Single vs Multiple Viewed Classification 90
5.3.4 Independent vs Dependent Classification 92
5.4 Variants of Text Categorization 94
5.4.1 Spam Mail Filtering 94
5.4.2 Sentimental Analysis 95
5.4.3 Information Filtering 97
5.4.4 Topic Routing 98
5.5 Summary and Further Discussions 99

6 Text Categorization: Approaches 101
6.1 Machine Learning 101
6.2 Lazy Learning 103
6.2.1 K Nearest Neighbor 104
6.2.2 Radius Nearest Neighbor 106
6.2.3 Distance-Based Nearest Neighbor 107
6.2.4 Attribute Discriminated Nearest Neighbor 109
6.3 Probabilistic Learning 110
6.3.1 Bayes Rule 111
6.3.2 Bayes Classifier 112
6.3.3 Naive Bayes 114
6.3.4 Bayesian Learning 116
6.4 Kernel Based Classifier 118
6.4.1 Perceptron 119
6.4.2 Kernel Functions 120
6.4.3 Support Vector Machine 122
6.4.4 Optimization Constraints 124
6.5 Summary and Further Discussions 126

7 Text Categorization: Implementation 129
7.1 System Architecture 129
7.2 Class Definitions 131
7.2.1 Classes: Word, Text, and PlainText 131
7.2.2 Interface and Class: Classifier and KNearestNeighbor 134
7.2.3 Class: TextClassificationAPI 136
7.3 Method Implementations 137
7.3.1 Class: Word 138
7.3.2 Class: PlainText 139
7.3.3 Class: KNearestNeighbor 141
7.3.4 Class: TextClassificationAPI 142
7.4 Graphic User Interface and Demonstration 145
7.4.1 Class: TextClassificationGUI 145
7.4.2 Preliminary Tasks and Encoding 147

7.4.3 Classification Process 152
7.4.4 System Upgrading 155
7.5 Summary and Further Discussions 156

8 Text Categorization: Evaluation 157
8.1 Evaluation Overview 157
8.2 Text Collections 159
8.2.1 NewsPage.com 159
8.2.2 20NewsGroups 160
8.2.3 Reuter21578 161
8.2.4 OSHUMED 163
8.3 F1 Measure 164
8.3.1 Contingency Table 165
8.3.2 Micro-Averaged F1 166
8.3.3 Macro-Averaged F1 168
8.3.4 Example 170
8.4 Statistical t-Test 171
8.4.1 Student's t-Distribution 171
8.4.2 Unpaired Difference Inference 174
8.4.3 Paired Difference Inference 175
8.4.4 Example 177
8.5 Summary and Further Discussions 178

Part III Text Clustering

9 Text Clustering: Conceptual View 183
9.1 Definition of Text Clustering 183
9.2 Data Clustering 184
9.2.1 SubSubsectionTitle 185
9.2.2 Association vs Clustering 186
9.2.3 Classification vs Clustering 187
9.2.4 Constraint Clustering 188
9.3 Clustering Types 189
9.3.1 Static vs Dynamic Clustering 190
9.3.2 Crisp vs Fuzzy Clustering 191
9.3.3 Flat vs Hierarchical Clustering 193
9.3.4 Single vs Multiple Viewed Clustering 194
9.4 Derived Tasks from Text Clustering 196
9.4.1 Cluster Naming 196
9.4.2 Subtext Clustering 197
9.4.3 Automatic Sampling for Text Categorization 199
9.4.4 Redundant Project Detection 200
9.5 Summary and Further Discussions 201

10 Text Clustering: Approaches 203
10.1 Unsupervised Learning 203
10.2 Simple Clustering Algorithms 204
10.2.1 AHC Algorithm 205
10.2.2 Divisive Clustering Algorithm 206
10.2.3 Single Pass Algorithm 207
10.2.4 Growing Algorithm 209
10.3 K Means Algorithm 211
10.3.1 Crisp K Means Algorithm 211
10.3.2 Fuzzy K Means Algorithm 212
10.3.3 Gaussian Mixture 213
10.3.4 K Medoid Algorithm 214
10.4 Competitive Learning 217
10.4.1 Kohonen Networks 217
10.4.2 Learning Vector Quantization 219
10.4.3 Two-Dimensional Self-Organizing Map 220
10.4.4 Neural Gas 222
10.5 Summary and Further Discussions 223

11 Text Clustering: Implementation 225
11.1 System Architecture 225
11.2 Class Definitions 227
11.2.1 Classes in Text Categorization System 227
11.2.2 Class: Cluster 230
11.2.3 Interface: ClusterAnalyzer 232
11.2.4 Class: AHCAlgorithm 233
11.3 Method Implementations 235
11.3.1 Methods in Previous Classes 235
11.3.2 Class: Cluster 237
11.3.3 Class: AHC Algorithm 239
11.4 Class: ClusterAnalysisAPI 240
11.4.1 Class: ClusterAnalysisAPI 241
11.4.2 Class: ClusterAnalyzerGUI 242
11.4.3 Demonstration 244
11.4.4 System Upgrading 245
11.5 Summary and Further Discussions 246

12 Text Clustering: Evaluation 249
12.1 Introduction 249
12.2 Cluster Validations 250
12.2.1 Intra-Cluster and Inter-Cluster Similarities 250
12.2.2 Internal Validation 252
12.2.3 Relative Validation 253
12.2.4 External Validation 255
12.3 Clustering Index 257
12.3.1 Computation Process 257

12.3.2 Evaluation of Crisp Clustering 258
12.3.3 Evaluation of Fuzzy Clustering 259
12.3.4 Evaluation of Hierarchical Clustering 261
12.4 Parameter Tuning 263
12.4.1 Clustering Index for Unlabeled Documents 263
12.4.2 Simple Clustering Algorithm with Parameter Tuning 264
12.4.3 K Means Algorithm with Parameter Tuning 265
12.4.4 Evolutionary Clustering Algorithm 266
12.5 Summary and Further Discussions 267

Part IV Advanced Topics

13 Text Summarization 271
13.1 Definition of Text Summarization 271
13.2 Text Summarization Types 272
13.2.1 Manual vs Automatic Text Summarization 273
13.2.2 Single vs Multiple Text Summarization 274
13.2.3 Flat vs Hierarchical Text Summarization 276
13.2.4 Abstraction vs Query-Based Summarization 278
13.3 Approaches to Text Summarization 279
13.3.1 Heuristic Approaches 280
13.3.2 Mapping into Classification Task 281
13.3.3 Sampling Schemes 283
13.3.4 Application of Machine Learning Algorithms 285
13.4 Combination with Other Text Mining Tasks 287
13.4.1 Summary-Based Classification 288
13.4.2 Summary-Based Clustering 289
13.4.3 Topic-Based Summarization 290
13.4.4 Text Expansion 292
13.5 Summary and Further Discussions 293

14 Text Segmentation 295
14.1 Definition of Text Segmentation 295
14.2 Text Segmentation Type 296
14.2.1 Spoken vs Written Text Segmentation 296
14.2.2 Ordered vs Unordered Text Segmentation 298
14.2.3 Exclusive vs Overlapping Segmentation 300
14.2.4 Flat vs Hierarchical Text Segmentation 302
14.3 Machine Learning-Based Approaches 304
14.3.1 Heuristic Approaches 304
14.3.2 Mapping into Classification 305
14.3.3 Encoding Adjacent Paragraph Pairs 307
14.3.4 Application of Machine Learning 309
14.4 Derived Tasks 311
14.4.1 Temporal Topic Analysis 311
14.4.2 Subtext Retrieval 313

14.4.3 Subtext Synthesization 314
14.4.4 Virtual Text 315
14.5 Summary and Further Discussions 316

15 Taxonomy Generation 319
15.1 Definition of Taxonomy Generation 319
15.2 Relevant Tasks to Taxonomy Generation 320
15.2.1 Keyword Extraction 320
15.2.2 Word Categorization 322
15.2.3 Word Clustering 324
15.2.4 Topic Routing 325
15.3 Taxonomy Generation Schemes 327
15.3.1 Index-Based Scheme 327
15.3.2 Clustering-Based Scheme 328
15.3.3 Association-Based Scheme 329
15.3.4 Link Analysis-Based Scheme 331
15.4 Taxonomy Governance 332
15.4.1 Taxonomy Maintenance 332
15.4.2 Taxonomy Growth 334
15.4.3 Taxonomy Integration 335
15.4.4 Ontology 337
15.5 Summary and Further Discussions 339

16 Dynamic Document Organization 341
16.1 Definition of Dynamic Document Organization 341
16.2 Online Clustering 342
16.2.1 Online Clustering in Functional View 342
16.2.2 Online K Means Algorithm 344
16.2.3 Online Unsupervised KNN Algorithm 345
16.2.4 Online Fuzzy Clustering 346
16.3 Dynamic Organization 348
16.3.1 Execution Process 348
16.3.2 Maintenance Mode 349
16.3.3 Creation Mode 350
16.3.4 Additional Tasks 351
16.4 Issues of Dynamic Document Organization 352
16.4.1 Text Representation 353
16.4.2 Binary Decomposition 353
16.4.3 Transition into Creation Mode 354
16.4.4 Variants of DDO System 355
16.5 Summary and Further Discussions 356

References 359

Index 365