

Texts in Computer Science

Editors

David Gries

Fred B. Schneider

For further volumes:

<http://www.springer.com/series/3191>

Sholom M. Weiss • Nitin Indurkha • Tong Zhang

Fundamentals of Predictive Text Mining

 Springer

Sholom M. Weiss
T.J. Watson Research Center
IBM Corporation
Kitchawan Road 1101
Yorktown Heights, 10598 NY
USA
sholom@data-miner.com

Tong Zhang
Dept. Statistics, Hill Center
Rutgers University
Piscataway, 08854-8019 NJ
USA
tongz@rci.rutgers.edu

Nitin Indurkha
School of Computer Science & Engg.
University of New South Wales
Sydney, 2052 NSW
Australia
nitin@data-miner.com

Series Editors

David Gries
Department of Computer Science
Upson Hall
Cornell University
Ithaca, NY 14853-7501, USA

Fred B. Schneider
Department of Computer Science
Upson Hall
Cornell University
Ithaca, NY 14853-7501, USA

ISSN 1868-0941
ISBN 978-1-84996-225-4
DOI 10.1007/978-1-84996-226-1
Springer London Dordrecht Heidelberg New York

e-ISSN 1868-095X
e-ISBN 978-1-84996-226-1

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010928348

© Springer-Verlag London Limited 2010

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Cover design: VTEX, Vilnius

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Five years ago, we authored “Text Mining: Predictive Methods for Analyzing Unstructured Information.” That book was geared mostly to professional practitioners, but was adaptable to course work with some effort by the instructor. Many topics were evolving, and this was one of the earliest efforts to collect material for predictive text mining. Since then, the book has seen extensive use in education, by ourselves and other instructors, with positive responses from students. With more data sourced from the Internet, the field has seen very rapid growth with many new techniques that would interest practitioners. Given the amount of supplementary new material we had begun using, a new edition was clearly needed. A year ago, our publisher asked us to update the book and to add material that would extend its use as a textbook. We have revised many sections, adding new material to reflect the increased use of the web. Exercises and summaries are also provided.

The prediction problem, looking for predictive patterns in data, has been widely studied. Strong methods are available to the practitioner. These methods process structured numerical information, where uniform measurements are taken over a sample of data. Text is often described as unstructured information. So, it would seem, text and numerical data are different, requiring different methods. Or are they? In our view, a prediction problem can be solved by the same methods, whether the data are structured numerical measurements or unstructured text. Text and documents can be transformed into measured values, such as the presence or absence of words, and the same methods that have proven successful for predictive data mining can be applied to text. Yet, there are key differences. Evaluation techniques must be adapted to the chronological order of publication and to alternative measures of error. Because the data are documents, more specialized analytical methods may be preferred for text. Moreover, the methods must be modified to accommodate very high dimensions: tens of thousands of words and documents. Still, the central themes are similar.

Our view of text mining allows us to unify the concepts of different fields. No longer is “natural language processing” the sole domain of linguists and their allied computer specialists. No longer is search engine technology distinct from other forms of machine learning. Ours is an open view. We welcome you to try your hand

at learning from data, whether numerical or text. Large text collections, often readily available on the Internet, contain valuable information that can be mined with today's tools instead of waiting for tomorrow's linguistic techniques. While others search for the essence of language understanding, we can immediately look for recurring word patterns in large collections of digital documents.

Our main theme is a strictly empirical view of text mining and an application of well-known analytical methods. We provide examples and software. Our presentation has a pragmatic bent with numerous references in the research literature for you to follow when so inclined. We want to be practical, yet inclusive of the wide community that might be interested in applications of text mining. We concentrate on predictive learning methods but also look at information retrieval and search engines, as well as clustering methods. We illustrate by examples, case studies, and the accompanying downloadable software.

While some analytical methods may be highly developed, predictive text mining is an emerging area of application. We have tried to summarize our experiences and provide the tools and techniques for your own experiments.

Audience

Our book is aimed at IT professionals and managers as well as advanced undergraduate computer science students and beginning graduate students. Some background in data mining is beneficial but is not essential. A few sections discuss advanced concepts that require mathematical maturity for a proper understanding. In such sections, intuitive explanations are also provided that may suffice for the less advanced reader. Most parts of the book can be read and understood by anyone with a sufficient analytic bend. If you are looking to do research in the area, the material in this book will provide direction in expanding your horizons. If you want to be a practitioner of text mining, you can read about our recommended methods and our descriptions of case studies. The software requires familiarity with running command-line programs and editing configuration files.

For Instructors

The material in this book has been successfully used for education in a variety of ways ranging from short intensive one-week courses to twelve-week full semester courses. In short courses, the mathematical material can be skipped. The exercises have undergone class-testing over several years. Each chapter has the following accompanying material:

- a chapter summary
- exercises.

In addition, numerous examples and figures are interlaced throughout the book, and these are available at data-miner.com as freely downloadable slides.

Supplementary Web Software

Data-Miner Pty. Ltd. has provided a free software license for those who have purchased the book. The software, which implements many of the methods discussed in the book, can be downloaded from the data-miner.com Web site. Linux scripts for many examples are also available for download. See the Appendix A for details.

Acknowledgements

Fred Damerau, our colleague and mentor, was a co-author of our original book. He is no longer with us, and his contributions to our project, especially his expertise in linguistics, were immeasurable. Some of the case studies in Chap. 7 are based on our prior publications. In those projects, we acknowledge the participation of Chidanand Apté, Radu Florian, Abraham Ittycheriah, Vijay Iyengar, Hongyan Jing, David Johnson, Frank Oles, Naval Verma, and Brian White. Arindam Banerjee made many helpful comments on a draft of our book. The exercises in the book evolved from our text-mining course conducted regularly at statistics.com. We thank our editor, Wayne Wheeler, and our previous editors Ann Kostant and Wayne Yuhasz, for their support.

New York, USA
Sydney, Australia

Sholom Weiss and Tong Zhang
Nitin Indurkha

Contents

1	Overview of Text Mining	1
1.1	What's Special About Text Mining?	1
1.1.1	Structured or Unstructured Data?	2
1.1.2	Is Text Different from Numbers?	3
1.2	What Types of Problems Can Be Solved?	5
1.3	Document Classification	6
1.4	Information Retrieval	6
1.5	Clustering and Organizing Documents	7
1.6	Information Extraction	8
1.7	Prediction and Evaluation	9
1.8	The Next Chapters	10
1.9	Summary	10
1.10	Historical and Bibliographical Remarks	11
1.11	Questions and Exercises	12
2	From Textual Information to Numerical Vectors	13
2.1	Collecting Documents	13
2.2	Document Standardization	15
2.3	Tokenization	16
2.4	Lemmatization	17
2.4.1	Inflectional Stemming	19
2.4.2	Stemming to a Root	19
2.5	Vector Generation for Prediction	21
2.5.1	Multiword Features	26
2.5.2	Labels for the Right Answers	28
2.5.3	Feature Selection by Attribute Ranking	29
2.6	Sentence Boundary Determination	29
2.7	Part-of-Speech Tagging	31
2.8	Word Sense Disambiguation	32
2.9	Phrase Recognition	32
2.10	Named Entity Recognition	33

2.11	Parsing	33
2.12	Feature Generation	35
2.13	Summary	36
2.14	Historical and Bibliographical Remarks	36
2.15	Questions and Exercises	38
3	Using Text for Prediction	39
3.1	Recognizing that Documents Fit a Pattern	41
3.2	How Many Documents Are Enough?	42
3.3	Document Classification	43
3.4	Learning to Predict from Text	44
3.4.1	Similarity and Nearest-Neighbor Methods	45
3.4.2	Document Similarity	46
3.4.3	Decision Rules	48
3.4.4	Decision Trees	54
3.4.5	Scoring by Probabilities	55
3.4.6	Linear Scoring Methods	58
3.5	Evaluation of Performance	66
3.5.1	Estimating Current and Future Performance	66
3.5.2	Getting the Most from a Learning Method	69
3.6	Applications	69
3.7	Summary	70
3.8	Historical and Bibliographical Remarks	70
3.9	Questions and Exercises	72
4	Information Retrieval and Text Mining	75
4.1	Is Information Retrieval a Form of Text Mining?	75
4.2	Key Word Search	76
4.3	Nearest-Neighbor Methods	77
4.4	Measuring Similarity	78
4.4.1	Shared Word Count	78
4.4.2	Word Count and Bonus	78
4.4.3	Cosine Similarity	79
4.5	Web-based Document Search	80
4.5.1	Link Analysis	81
4.6	Document Matching	85
4.7	Inverted Lists	85
4.8	Evaluation of Performance	87
4.9	Summary	88
4.10	Historical and Bibliographical Remarks	88
4.11	Questions and Exercises	89
5	Finding Structure in a Document Collection	91
5.1	Clustering Documents by Similarity	93
5.2	Similarity of Composite Documents	94
5.2.1	k -Means Clustering	96

5.2.2	Hierarchical Clustering	99
5.2.3	The EM Algorithm	102
5.3	What Do a Cluster's Labels Mean?	105
5.4	Applications	107
5.5	Evaluation of Performance	108
5.6	Summary	110
5.7	Historical and Bibliographical Remarks	110
5.8	Questions and Exercises	111
6	Looking for Information in Documents	113
6.1	Goals of Information Extraction	113
6.2	Finding Patterns and Entities from Text	115
6.2.1	Entity Extraction as Sequential Tagging	116
6.2.2	Tag Prediction as Classification	117
6.2.3	The Maximum Entropy Method	118
6.2.4	Linguistic Features and Encoding	123
6.2.5	Local Sequence Prediction Models	124
6.2.6	Global Sequence Prediction Models	128
6.3	Coreference and Relationship Extraction	129
6.3.1	Coreference Resolution	129
6.3.2	Relationship Extraction	131
6.4	Template Filling and Database Construction	132
6.5	Applications	133
6.5.1	Information Retrieval	133
6.5.2	Commercial Extraction Systems	134
6.5.3	Criminal Justice	135
6.5.4	Intelligence	135
6.6	Summary	136
6.7	Historical and Bibliographical Remarks	137
6.8	Questions and Exercises	138
7	Data Sources for Prediction: Databases, Hybrid Data and the Web	141
7.1	Ideal Models of Data	141
7.1.1	Ideal Data for Prediction	141
7.1.2	Ideal Data for Text and Unstructured Data	142
7.1.3	Hybrid and Mixed Data	142
7.2	Practical Data Sourcing	144
7.3	Prototypical Examples	145
7.3.1	Web-based Spreadsheet Data	146
7.3.2	Web-based XML Data	146
7.3.3	Opinion Data and Sentiment Analysis	148
7.4	Hybrid Example: Independent Sources of Numerical and Text Data	151
7.5	Mixed Data in Standard Table Format	152
7.6	Summary	153
7.7	Historical and Bibliographical Remarks	154
7.8	Questions and Exercises	154

8	Case Studies	157
8.1	Market Intelligence from the Web	157
8.1.1	The Problem	157
8.1.2	Solution Overview	158
8.1.3	Methods and Procedures	159
8.1.4	System Deployment	160
8.2	Lightweight Document Matching for Digital Libraries	161
8.2.1	The Problem	161
8.2.2	Solution Overview	162
8.2.3	Methods and Procedures	163
8.2.4	System Deployment	164
8.3	Generating Model Cases for Help Desk Applications	165
8.3.1	The Problem	165
8.3.2	Solution Overview	165
8.3.3	Methods and Procedures	166
8.3.4	System Deployment	168
8.4	Assigning Topics to News Articles	169
8.4.1	The Problem	169
8.4.2	Solution Overview	169
8.4.3	Methods and Procedures	169
8.4.4	System Deployment	173
8.5	E-mail Filtering	174
8.5.1	The Problem	174
8.5.2	Solution Overview	174
8.5.3	Methods and Procedures	175
8.5.4	System Deployment	177
8.6	Search Engines	177
8.6.1	The Problem	177
8.6.2	Solution Overview	177
8.6.3	Methods and Procedures	178
8.6.4	System Deployment	179
8.7	Extracting Named Entities from Documents	181
8.7.1	The Problem	181
8.7.2	Solution Overview	181
8.7.3	Methods and Procedures	182
8.7.4	System Deployment	184
8.8	Customized Newspapers	184
8.8.1	The Problem	184
8.8.2	Solution Overview	185
8.8.3	Methods and Procedures	186
8.8.4	System Deployment	187
8.9	Summary	187
8.10	Historical and Bibliographical Remarks	188
8.11	Questions and Exercises	188

9	Emerging Directions	189
9.1	Summarization	189
9.2	Active Learning	192
9.3	Learning with Unlabeled Data	193
9.4	Different Ways of Collecting Samples	194
9.4.1	Ensembles and Voting Methods	194
9.4.2	Online Learning	196
9.4.3	Cost-Sensitive Learning	197
9.4.4	Unbalanced Samples and Rare Events	198
9.5	Distributed Text Mining	198
9.6	Learning to Rank	200
9.7	Question Answering	201
9.8	Summary	202
9.9	Historical and Bibliographical Remarks	203
9.10	Questions and Exercises	204
A	Software Notes	207
A.1	Summary of Software	207
A.2	Requirements	208
A.3	Download Instructions	208
	References	211
	Author Index	219
	Subject Index	223