SOCIAL MEDIA MINING

The growth of social media over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social media. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research, novel algorithms, and tool development.

Social Media Mining integrates social media, social network analysis, and data mining to provide a convenient and coherent platform for students, practitioners, researchers, and project managers to understand the basics and potentials of social media mining. It introduces the unique problems arising from social media data and presents fundamental concepts, emerging issues, and effective algorithms for network analysis and data mining.

Suitable for use in advanced undergraduate and beginning graduate courses as well as professional short courses, the text contains exercises of different degrees of difficulty that improve understanding and help apply concepts, principles, and methods in various scenarios of social media mining.

Reza Zafarani is a research associate of Computer Science and Engineering at Arizona State University. His research interests are in social media mining, machine learning, social network analysis, and social computing. His research emphasis is on user behavioral analysis at scale, and information integration and modeling across social media sites.

Mohammad Ali Abbasi is a research associate of Computer Science and Engineering at Arizona State University. His research interests are in data mining, machine learning, and social computing. In particular, his research is focused on user profiling, user credibility assessment, and real-world applications of social media.

Huan Liu is a professor of Computer Science and Engineering at Arizona State University where he has been recognized for excellence in teaching and research. His research interests are in data mining, machine learning, social computing, artificial intelligence, and investigating problems that arise in real-world data-intensive applications with high-dimensional data of disparate forms, such as social media.

SOCIAL MEDIA MINING

An Introduction

REZA ZAFARANI
Arizona State University, Tempe

MOHAMMAD ALI ABBASI
Arizona State University, Tempe

HUAN LIU
Arizona State University, Tempe

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107018853

First published 2014

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data
Zafarani, Reza, 1983–
Social media mining : an introduction / Reza Zafarani, Arizona State University, Tempe, Mohammad Ali Abbasi, Arizona State University, Tempe, Huan Liu, Arizona State University, Tempe.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-01885-3 (hardback)
1. Data mining. 2. Social media – Research. 3. Behavioral assessment – Data processing. 4. Webometrics. I. Abbasi, Mohammad Ali, 1975– II. Liu, Huan, 1958– III. Title.
QA76.9.D343Z34 2014
006.3′12–dc23 2013035271

ISBN 978-1-107-01885-3 Hardback

Additional resources for this publication at http://dmml.asu.edu/smm

To our families . . .

Contents

Preface

We live in an age of big data. With hundreds of millions of people spending countless hours on social media to share, communicate, connect, interact, and create user-generated data at an unprecedented rate, social media has become one unique source of big data. This novel source of rich data provides unparalleled opportunities and great potential for research and development. Unfortunately, more data does not necessarily beget more good, only more of the right (or relevant) data that enables us to glean gems. Social media data differs from traditional data we are familiar with in data mining. Thus, new computational methods are needed to mine the data. Social media data is noisy, free-format, of varying length, and multimedia. Furthermore, social relations among the entities, or social networks, form an inseparable part of social media data; hence, it is important that social theories and research methods be employed with statistical and data mining methods. It is therefore a propitious time for social media mining.

Social media mining is a rapidly growing new field. It is an interdisciplinary field at the crossroad of disparate disciplines deeply rooted in computer science and social sciences. There are an active community and a large body of literature about social media. The fast-growing interests and intensifying need to harness social media data require research and the development of tools for finding insights from big social media data. This book is one of the intellectual efforts to answer the novel challenges of social media. It is designed to enable students, researchers, and practitioners to acquire fundamental concepts and algorithms for social media mining.

Researchers in this emerging field are expected to have knowledge in different areas, such as data mining, machine learning, text mining, social network analysis, and information retrieval, and are often required to consult research papers to learn the state of the art of social media mining. To mitigate such a strenuous effort and help researchers get up to speed in a

convenient way, we take advantage of our teaching and research of many years to survey, summarize, filter, categorize, and connect disparate research findings and fundamental concepts of social media mining. This book is our diligent attempt to provide an easy reference or entry point to help researchers quickly acquire a wide range of essentials of social media mining. Social media not only produces big user-generated data; it also has a huge potential for social science research, business development, and understanding human and group behavior. If you want to share a piece of information or a site on social media, you would like to grab precious attention from other equally eager users of social media; if you are curious to know what is hidden or who is influential in the complex world of social media, you might wonder how one can find this information in big and messy social media; if you hope to serve your customers better in social media, you certainly want to employ effective means to understand them better. These are just some scenarios in which social media mining can help. If one of these scenarios fits your need or you simply wish to learn something interesting in this emerging field of social media mining, this book is for you. We hope this book can be of benefit to you in accomplishing your goals of dealing with big data of social media.

Book Website and Resources

The book's website and further resources can be found at

`http://dmml.asu.edu/smm`

The website provides lecture slides, homework and exam problems, and sample projects, as well as pointers to useful material and resources that are publicly available and relevant to social media mining.

To Instructors

The book is designed for a one-semester course for senior undergraduate or graduate students. Though it is mainly written for students with a background in computer science, readers with a basic understanding of probability, statistics, and linear algebra will find it easily accessible. Some chapters can be skipped or assigned as a homework assignment for reviewing purposes if students have knowledge of a chapter. For example, if students have taken a data mining or machine learning course, they can skip Chapter 5. When time is limited, Chapters 6–8 should be discussed in

depth, and Chapters 9 and 10 can be either discussed briefly or assigned as part of reading material for course projects.

Reza Zafarani
Mohammad Ali Abbasi
Huan Liu
Tempe, AZ
August 2013

Acknowledgments

In the past several years, enormous pioneering research has been performed by numerous researchers in the interdisciplinary fields of data mining, social computing, social network analysis, network science, computer science, and the social sciences. We are truly dwarfed by the depth, breadth, and extent of the literature, which not only made it possible for us to complete a text on this emerging topic – *social media mining* – but also made it a seemingly endless task. In the process, we have been fortunate in drawing inspiration and obtaining great support and help from many people to whom we are indebted.

We would like to express our tremendous gratitude to the current and former members of the Data Mining and Machine Learning laboratory at Arizona State University (ASU); in particular, Nitin Agrawal, Salem Alelyani, Geoffrey Barbier, William Cole, Zhuo Feng, Magdiel Galan-Oliveras, Huiji Gao, Pritam Gundecha, Xia (Ben) Hu, Isaac Jones, Shamanth Kumar, Fred Morstatter, Sai Thejasvee Moturu, Ashwin Rajadesingan, Suhas Ranganath, Jiliang Tang, Lei Tang, Xufei Wang, and Zheng Zhao. Without their impressive accomplishments and continuing strides in advancing research in data mining, machine learning, and social computing, this book would have not been possible. Their stimulating thoughts, creative ideas, friendly aggressiveness, willingness to extend the research frontier, and cool company during our struggling moments (Arizona could be scorchingly hot in some months), directly and indirectly, offered us encouragement, drive, passion, and ideas, as well as critiques in the process toward the completion of the book.

This book project stemmed from a course on social computing offered in 2008 at ASU. It was a seminar course that enjoyed active participation by graduate students and bright undergraduates with intelligent and provocative minds. Lively discussion and heated arguments were fixtures of the seminar course. Since then, it has become a regular course, evolving into a focused theme on *social media mining*. Teaching assistants, students, and guest speakers in these annual courses were of significant help to us in

choosing topics to include, determining the depth and extent of each topic, and offering feedback on lecture materials such as homework problems, slides, course projects, and reading materials.

We would like to especially thank Denny Abraham Cheriyan, Nitin Ahuja, Amy Baldwin, Sai Prasanna Baskaran, Gaurav Pandey, Prerna Satija, Nitesh Kedia, Bernard Ma, Dhanyatha Manjunath, Girish Kumar Reddy Marthala, Apurv Patki, Greggory Scherer, Nikhil Sunkesula Bhaskar, Yajin Wang, and Ruozhou Yu for their detailed comments on the drafts of this book. In addition, we owe our gratitude to Daria Bazzi and Michael Meeder for their help in proofreading earlier versions of the book, Farzad Zafarani for preparing the book's website, Subbarao Kambhampati for reading an earlier draft and offering encouragement, and Rebecca Goolsby for continually challenging us in understanding social media and developing social computing tools for real-world applications of humanitarian assistance and disaster relief.

The idea of having this book published by Cambridge University Press began with a casual conversation with K. Selcuk Candan, a well-received Cambridge author. It was an enjoyable and pleasant process working with Cambridge University Press. We would like to thank Lauren Cowles, a senior editor of Mathematics and Computer Sciences at Cambridge, for her patience and kind support during the process, and the Cambridge team, David Jou and Joshua Penney, as well as Adrian Pereira and his colleagues at Aptara for their work on the production of the book.

Our research on social computing, data mining, and social network analysis has been, in part, supported by the Office of Naval Research, National Science Foundation, and Army Research Office.

We have truly enjoyed our collaboration in this arduous journey. We will certainly miss our weekly meetings and many missed deadlines.