Community Detection and Mining in Social Media

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, and Johannes Gehrke. 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. The scope will largely follow the purview of premier computer science conferences, such as KDD. Potential topics include, but not limited to, 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.

Community Detection and Mining in Social Media

Lei Tang and Huan Liu

ISBN: 978-3-031-00772-9 paperback
ISBN: 978-3-031-01900-5 ebook

DOI 10.1007/978-3-031-01900-5

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

Lecture #3
Series Editor: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of Maryland*
Wei Wang, *University of North Carolina, Chapel Hill*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Illinois, Chicago*
Series ISSN
Synthesis Lectures on Data Mining and Knowledge Discovery
Print 2151-0067 Electronic 2151-0075

Community Detection and Mining in Social Media

Lei Tang
Yahoo! Labs

Huan Liu
Arizona State University

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY
#3

ABSTRACT

The past decade has witnessed the emergence of participatory Web and social media, bringing people together in many creative ways. Millions of users are playing, tagging, working, and socializing online, demonstrating new forms of collaboration, communication, and intelligence that were hardly imaginable just a short time ago. Social media also helps reshape business models, sway opinions and emotions, and opens up numerous possibilities to study human interaction and collective behavior in an unparalleled scale. This lecture, from a data mining perspective, introduces characteristics of social media, reviews representative tasks of computing with social media, and illustrates associated challenges. It introduces basic concepts, presents state-of-the-art algorithms with easy-to-understand examples, and recommends effective evaluation methods. In particular, we discuss graph-based community detection techniques and many important extensions that handle dynamic, heterogeneous networks in social media. We also demonstrate how discovered patterns of communities can be used for social media mining. The concepts, algorithms, and methods presented in this lecture can help harness the power of social media and support building socially-intelligent systems. This book is an accessible introduction to the study of *community detection and mining in social media*. It is an essential reading for students, researchers, and practitioners in disciplines and applications where social media is a key source of data that piques our curiosity to understand, manage, innovate, and excel.

This book is supported by additional materials, including lecture slides, the complete set of figures, key references, some toy data sets used in the book, and the source code of representative algorithms. The readers are encouraged to visit the book website for the latest information:

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

KEYWORDS

social media, community detection, social media mining, centrality analysis, strength of ties, influence modeling, information diffusion, influence maximization, correlation, homophily, influence, community evaluation, heterogeneous networks, multi-dimensional networks, multi-mode networks, community evolution, collective classification, social dimension, behavioral study

To my parents and wife — LT

To my parents, wife, and sons — HL

Contents

Acknowledgments

It is a pleasure to acknowledge many colleagues who made substantial contributions in various ways to this time-consuming book project. The members of the Social Computing Group, Data Mining and Machine Learning Lab at Arizona State University made this project enjoyable. They include Ali Abbasi, Geoffrey Barbier, William Cole, Gabriel Fung, Huiji Gao, Shamanth Kumar, Xufei Wang, and Reza Zafarani. Particular thanks go to Reza Zafarani and Gabriel Fung who read the earlier drafts of the manuscript and provided helpful comments to improve the readability.

We are grateful to Professor Sun-Ki Chai, Professor Michael Hechter, Dr. John Salerno and Dr. Jianping Zhang for many inspiring discussions. This work is part of the projects sponsored by grants from AFOSR and ONR.

We thank Morgan & Claypool and particularly executive editor Diane D. Cerra for her help and patience throughout this project. We also thank Professor Nitin Agarwal at University of Arkansas at Little Rock for his prompt answers to questions concerning the book editing.

Last and foremost, we thank our families for supporting us through this fun project. We dedicate this book to them, with love.

Lei Tang and Huan Liu
August, 2010