This chapter is from *Social Media Mining: An Introduction*.
By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu.
Cambridge University Press, 2014. Draft version: April 20, 2014.
Complete Draft and Slides Available at: http://dmml.asu.edu/smm

Chapter 1

Introduction

With the rise of *social media*, the web has become a vibrant and lively realm in which billions of individuals all around the globe interact, share, post, and conduct numerous daily activities. Information is collected, curated, and published by *citizen journalists* and simultaneously shared or consumed by thousands of individuals, who give spontaneous feedback. Social media enables us to be connected and interact with each other anywhere and anytime – allowing us to observe human behavior in an unprecedented scale with a new lens. This social media lens provides us with golden opportunities to understand individuals at scale and to mine human behavioral patterns otherwise impossible. As a byproduct, by understanding individuals better, we can design better computing systems tailored to individuals' needs that will serve them and society better. This new social media world has no geographical boundaries and incessantly churns out oceans of data. As a result, we are facing an exacerbated problem of big data – "drowning in data, but thirsty for knowledge." Can data mining come to the rescue?

Social Media

Citizen Journalism

Unfortunately, social media data is significantly different from the traditional data that we are familiar with in data mining. Apart from enormous size, the mainly user-generated data is noisy and unstructured, with abundant social relations such as friendships and followers-followees. This new type of data mandates new computational data analysis approaches that can combine social theories with statistical and data mining methods. The pressing demand for new techniques ushers in and entails a new interdisciplinary field – social media mining.

1.1 What is Social Media Mining

Social media shatters the boundaries between the real world and the virtual world. We can now integrate social theories with computational methods to study how individuals (also known as *social atoms*) interact and how communities (i.e., *social molecules*) form. The uniqueness of social media data calls for novel data mining techniques that can effectively handle user-generated content with rich social relations. The study and development of these new techniques are under the purview of social media mining, an emerging discipline under the umbrella of data mining. ***Social Media Mining*** *is the process of representing, analyzing, and extracting actionable patterns from social media data*.

Social Atom

Social Molecule

Social Media Mining

Social Media Mining, introduces basic concepts and principal algorithms suitable for investigating massive social media data; it discusses theories and methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and mathematics. It encompasses the tools to formally represent, measure, model, and mine meaningful patterns from large-scale social media data.

Data Scientist

Social media mining cultivates a new kind of *data scientist* who is well versed in social and computational theories, specialized to analyze recalcitrant social media data, and skilled to help bridge the gap from what we know (social and computational theories) to what we want to know about the vast social media world with computational tools.

1.2 New Challenges for Mining

Social media mining is an emerging field where there are more problems than ready solutions. Equipped with interdisciplinary concepts and theories, fundamental principles, and state-of-the-art algorithms, we can stand on the shoulders of the giants and embark on solving challenging problems and developing novel data mining techniques and scalable computational algorithms. In general, social media can be considered a world of social atoms (i.e., individuals), entities (e.g., content, sites, networks, etc.), and interactions between individuals and entities. Social theories and social norms govern the interactions between individuals and entities. For effective social media mining, we collect information about individuals and

entities, measure their interactions, and discover patterns to understand human behavior.

Mining social media data is the task of mining user-generated content with social relations. This data[1] presents novel challenges encountered in social media mining.

Big Data Paradox

Big Data Paradox. Social media data is undoubtedly big. However, when we zoom into individuals for whom, for example, we would like to make relevant recommendations, we often have little data for each specific individual. We have to exploit the characteristics of social media and use its multidimensional, multisource, and multisite data to aggregate information with sufficient statistics for effective mining.

Obtaining Sufficient Samples

Obtaining Sufficient Samples. One of the commonly used methods to collect data is via application programming interfaces (APIs) from social media sites. Only a limited amount of data can be obtained daily. Without knowing the population's distribution, how can we know that our samples are reliable representatives of the full data? Consequently, how can we ensure that our findings obtained from social media mining are any indication of true patterns that can benefit our research or business development?

Noise Removal Fallacy

Noise Removal Fallacy. In classic data mining literature, a successful data mining exercise entails extensive data preprocessing and noise removal as "garbage in and garbage out." By its nature, social media data can contain a large portion of noisy data. For this data, we notice two important observations: (1) blindly removing noise can worsen the problem stated in the big data paradox because the removal can also eliminate valuable information, and (2) the definition of noise becomes complicated and relative because it is dependent on our task at hand.

Evaluation Dilemma

Evaluation Dilemma. A standard procedure of evaluating patterns in data mining is to have some kind of ground truth. For example, a dataset can be divided into training and test sets. Only the training data is used in learning, and the test data serves as ground truth for testing. However, ground truth is often not available in social media mining. Evaluating patterns from social media mining poses a seemingly insurmountable challenge. On the other hand, without credible evaluation, how can we guarantee the

[1]The data has a power-law distribution and more often than not, data is not independent and identically distributed (i.i.d.) as generally assumed in data mining.

validity of the patterns?

This book contains basic concepts and fundamental principles that will help readers contemplate and design solutions to address these challenges intrinsic to social media mining.

1.3 Book Overview and Reader's Guide

This book consists of three parts. Part I, *Essentials*, outlines ways to represent social media data and provides an understanding of fundamental elements of social media mining. Part II, *Communities and Interactions*, discusses how communities can be found in social media and how interactions occur and information propagates in social media. Part III, *Applications*, offers some novel illustrative applications of social media mining. Throughout the book, we use examples to explain how things work and to deepen the understanding of abstract concepts and profound algorithms. These examples show in a tangible way how theories are applied or ideas are materialized in discovering meaningful patterns in social media data.

Consider an online social networking site with millions of members in which members have the opportunity to befriend one another, send messages to each other, and post content on the site. Facebook, LinkedIn, and Twitter are exemplars of such sites. To make sense of data from these sites, we resort to social media mining to answer corresponding questions. In Part I: Essentials (Chapters 2–5), we learn to answer questions such as the following:

1. Who are the most important people in a social network?
2. How do people befriend others?
3. How can we find interesting patterns in user-generated content?

These essentials come into play in Part II: Communities and Interactions (Chapters 6 and 7) where we attempt to analyze how communities are formed, how they evolve, and how the qualities of detected communities are evaluated. We show ways in which information diffusion in social media can be studied. We aim to answer general questions such as the following:

1. How can we identify communities in a social network?

2. When someone posts an interesting article on a social network, how far can the article be transmitted in that network?

In Part III: Application (Chapters 8–10), we exemplify social media mining using real-world problems in dealing with social media: measuring influence, recommending in a social environment, and analyzing user behavior. We aim to answer these questions:

1. How can we measure the influence of individuals in a social network?
2. How can we recommend content or friends to individuals online?
3. How can we analyze the behavior of individuals online?

To provide an overall picture of the book content, we created a dependency graph among chapters (Fig. 1.1) in which arrows suggest dependencies between chapters. Based on the dependency graph, therefore, a reader can start with Chapter 2 (graph essentials), and it is recommended that he or she read Chapters 5 (data mining essentials) and 8 (influence and homophily) before Chapter 9 (recommendation in social media). We have also color-coded chapter boxes that are of the same level of importance and abstraction. The darkest chapters are the essentials of this book, and the lightest boxes are those chapters that are more applied and have materials that are built on the foundation of other chapters.

Who Should Read This Book?

A reader with a basic computer science background and knowledge of data structures, search, and graph algorithms will find this book easily accessible. Limited knowledge of linear algebra, calculus, probability, and statistics will help readers understand technical details with ease. Having a data mining or machine learning background is a plus, but not necessary.

The book is designed for senior undergraduate and graduate students. It is organized in such a way that it can be taught in one semester to students with a basic prior knowledge of statistics and linear algebra. It can also be used for a graduate seminar course by focusing on more advanced chapters with the supplement of detailed bibliographical notes. Moreover, the book can be used as a reference book for researchers, practitioners, and project managers of related fields who are interested in both learning the

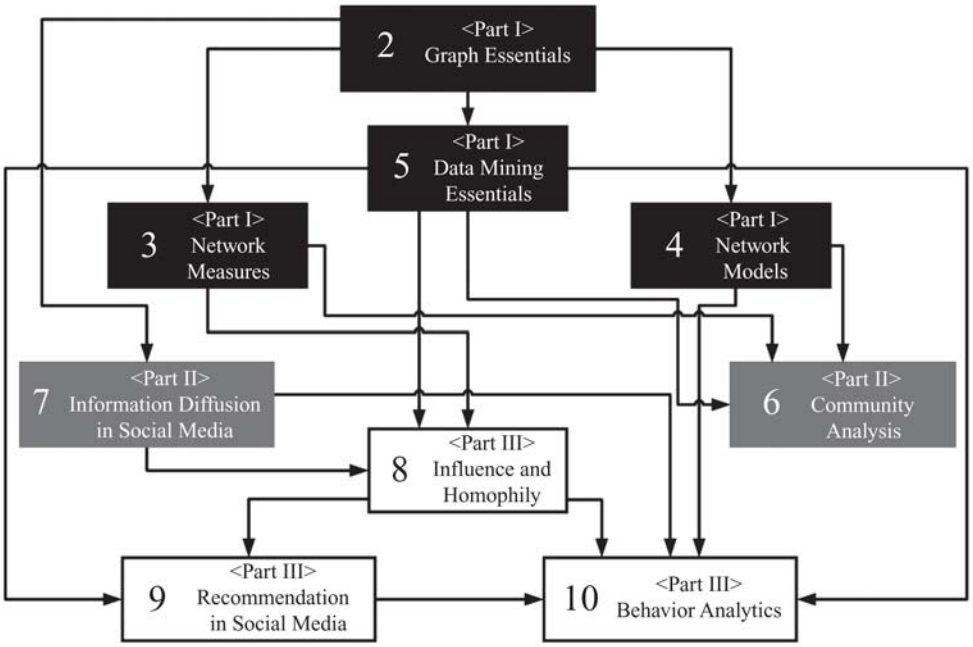

Figure 1.1: Dependency between Book Chapters. Arrows show dependencies and colors represent book parts.

basics and tangible examples of this emerging field and understanding the potentials and opportunities that social media mining can offer.

1.4 Summary

As defined by Kaplan and Haenlein [141], social media is the "group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content." There are many categories of social media including, but not limited to, social networking (Facebook or LinkedIn), microblogging (Twitter), photo sharing (Flickr, Photobucket, or Picasa), news aggregation (Google reader, StumbleUpon, or Feedburner), video sharing (YouTube, MetaCafe), livecasting (Ustream or Justin.TV), virtual worlds (Kaneva), social gaming (World of Warcraft), social search (Google, Bing, or Ask.com), and instant messaging (Google Talk, Skype, or Yahoo! messenger).

The first social media site was introduced by Geocities in 1994, which allowed users to create their own homepages. The first social networking site, SixDegree.com, was introduced in 1997. Since then, many other social media sites have been introduced, each providing service to millions of people. These individuals form a virtual world in which individuals (social atoms), entities (content, sites, etc.) and interactions (between individuals, between entities, between individuals and entities) coexist. Social norms and human behavior govern this virtual world. By understanding these social norms and models of human behavior and combining them with the observations and measurements of this virtual world, one can systematically analyze and mine social media.

Social media mining is the process of representing, analyzing, and extracting meaningful patterns from data in social media, resulting from social interactions. It is an interdisciplinary field encompassing techniques from computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and mathematics. Social media mining faces grand challenges such as the big data paradox, obtaining sufficient samples, the noise removal fallacy, and evaluation dilemma.

Social media mining represents the virtual world of social media in a computable way, measures it, and designs models that can help us understand its interactions. In addition, social media mining provides necessary tools to mine this world for interesting patterns, analyze information diffusion, study influence and homophily, provide effective recommendations, and analyze novel social behavior in social media.

1.5 Bibliographic Notes

For historical notes on social media sites and challenges in social media, refer to [81, 173, 141, 150, 115]. Kaplan and Haenlein [141] provide a categorization of social media sites into collaborative projects, blogs, content communities, social networking sites, virtual game worlds, and virtual social worlds. Our definition of social media is a rather abstract one whose elements are social atoms (individuals), entities, and interactions. A more detailed abstraction can be found in the work of [149]. They consider the seven building blocks of social media to be identity, conversation, sharing, presence, relationships, reputation, and groups. They argue that the amount of attention that sites give to these building blocks makes them different in nature. For instance, YouTube provides more functionality in terms of groups than LinkedIn.

Social media mining brings together techniques from many disciplines. General references that can accompany this book and help readers better understand the material in this book can be found in data mining and web mining [120, 280, 92, 174, 51], machine learning [40], and pattern recognition [75] texts, as well as network science and social network analysis [78, 253, 212, 140, 28] textbooks. For relevant references on optimization refer to [44, 219, 228, 207] and for algorithms to [61, 151]. For general references on social research methods consult [36, 47]. Note that these are generic references and more specific references are provided at the end of each chapter. This book discusses non–multimedia data in social media. For multimedia data analysis refer to [49].

Recent developments in social media mining can be found in journal articles in IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Knowledge Discovery from Data (TKDD), ACM Transactions on Intelligent Systems and Technology (TIST), Social Network Analysis and Mining (SNAM), Knowledge and Information Systems (KAIS), ACM Transactions on the Web (TWEB), Data Mining and Knowledge Discovery (DMKD), World Wide Web Journal, Social Networks, Internet Mathematics, IEEE Intelligent Systems, and SIGKDD Exploration. Conference papers can be found in proceedings of Knowledge Discovery and Data Mining (KDD), World Wide Web (WWW), Association for Computational Linguistics (ACL), Conference on Information and Knowledge Management (CIKM), International Conference on Data Mining (ICDM), Internet Measuring Conference (IMC), International Con-

ference on Weblogs and Social Media (ICWSM), International Conference on Web Engineering (ICWE), Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Data Basis (ECML/PKDD), Web Search and Data Mining (WSDM), International Joint Conferences on Artificial Intelligence (IJCAI), Association for the Advancement of Artificial Intelligence (AAAI), Recommender Systems (RecSys), Computer-Human Interaction (CHI), SIAM International Conference on Data Mining (SDM), Hypertext (HT), and Social Computing Behavioral-Cultural Modeling and Prediction (SBP) conferences.

Table 1.1: List of Websites

Amazon	Flickr	Facebook	Twitter
BlogCatalog	MySpace	Last.fm	Pandora
LinkedIn	Reddit	Vimeo	Del.icio.us
StumbleUpon	Yelp	YouTube	Meetup

1.6 Exercises

1. Discuss some methodologies that can address the grand challenges of social media.

2. What are the key characteristics of social media that differentiate it from other media? Please list at least two with a brief explanation.

3. What are the different types of social media? Name two, and provide a definition and an example for each type.

4. (a) Visit the websites in Table 1.1 (or find similar ones) and identify the types of activities that individuals can perform on each one.

   (b) Similar to questions posed in Section 1.3, design two questions that you find interesting to ask with respect to each site.

5. What marketing opportunities do you think exist in social media? Can you outline an example of such an opportunity in Twitter?

6. How does behavior of individuals change across sites? What behaviors remain consistent and what behaviors likely change? What are possible reasons behind these differences?

7. How does social media influence real-world behaviors of individuals? Identify a behavior that is due to the usage of, say, Twitter.

8. Outline how social media can help NGOs fulfill their missions better in performing tasks such as humanitarian assistance and disaster relief.

9. Identify at least three major side effects of information sharing on social media.

10. Rumors spread rapidly on social media. Can you think of some method to block the spread of rumors on social media?