Xun Liang

Social Computing with Artificial Intelligence

清華大學出版社
TSINGHUA UNIVERSITY PRESS

Springer

Social Computing with Artificial Intelligence

Xun Liang

Social Computing with Artificial Intelligence

清華大學出版社
TSINGHUA UNIVERSITY PRESS

Springer

Xun Liang
School of Information
Renmin University of China
Beijing, China

ISBN 978-981-15-7759-8 ISBN 978-981-15-7760-4 (eBook)
https://doi.org/10.1007/978-981-15-7760-4

Jointly published with Tsinghua University Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: Tsinghua University Press, Beijing, China.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Over recent years, the social network platform has seen unprecedented rapid development. These network platforms (such as Facebook, Twitter, Sina Weibo, Wechat, Netease News) not only gather a large number of users, but also provide huge potential user resources. These user resources promote the production of massive data sets in the field of business and scientific research, which are of vital significance for analyzing social group behavior and exploring potential huge social value. Therefore, a new discipline emerged at the historic moment—social computing.

As a new interdisciplinary research field, social computing, there is no recognized definition at present. However, we can analyze concepts from the background of social computing and generalize social computing as "using social methods to calculate society." The so-called socialization method is a method that centers on "grass-roots" users and relies on "grass-roots" users. It is a method of synergy and swarm intelligence, a mode of thinking from individual to whole, from micro to macro. Many incidents have been developed into a major social event by countless netizens' words and insignificant microbehavior. From this point of view, social computing is a computational model of swarm intelligence. It can be seen that the essence of social computing is the process of intelligent analysis of large social network data.

The research object of social computing is society, including physical society and virtual network society. The former mainly refers to our traditional sense of society, such as a country or region; the latter mainly refers to the web-based virtual network community. Virtual network society is a reflection of real physical society. The ultimate goal of studying virtual network society is to serve the management of real physical society. The study of social network activities will help to find the stable network that has remained for a long time, excavate the huge potential value, and provide important help for decision-makers to make more effective decision analysis.

The study of social network cannot be separated from the collision with artificial intelligence or machine learning, which is the most active branch of computer science at present. In this book, we also introduce supervised, unsupervised, and semi-supervised learning models as well as more specific state-of-the-art artificial intelligence algorithms, such as deep learning, reinforcement learning, brother learning, and epiphany learning. As a field leading the revolutionary change of information technology, the rise of artificial intelligence has promoted the rapid development of the Internet and the wide application of social networks, making the dissemination of information faster and the fermentation of public opinion warmer. Therefore, public opinion also needs modern means to manage. Public opinion is a kind of public composed of individuals and various social groups. In a certain historical stage and social space, it is the sum of various emotions, willingness, attitudes, and opinions held by various public firms which are closely related to their own interests. Network is the carrier of public opinion dissemination in modern society and has given new characteristics to the dissemination of public opinion.

Social network public opinion is a specific form of social public opinion, and a collection of common opinions which have certain influence and tendencies on certain phenomena, problems, or specific things publicly expressed by the public on the social network. However, the social network public opinion further narrows the network public opinion, which is limited to the relevant public opinion content of this subject. In recent years, public opinion monitoring has become a necessary technology for the state to manage the Internet. Using computer intelligence technology to transform various human emotions into real numerical data has become a research hotspot.

This book consists of twelve chapters, starting with low-level data processing, and gradually transiting to higher-level social computing and its applications. It is divided into three parts: Part I, data (including Chaps. 2 and 3), introduces the social computing data collection and data analysis algorithm, which correspond to social computing in the four areas of social data perception, i.e., knowledge discovery, individual and group community modeling, information dissemination and network dynamic evolution. Part II, models (including Chaps. 4–8), introduces the social computing model in three aspects: online crowd opinion content mining, community network structure, dynamic information dissemination, network propagation mechanism, which correspond to the areas of social computing. Part III, applications (including Chaps. 9–12), introduces the application cases of decision-making in the fields of social computing to support social practical problems.

In order to help readers have a stronger appreciation of the concepts, the book contains a large number of data, charts and references, including data and algorithms that we think are vital, available and explanatory, so that readers can dig into social computing through this book as much as possible.

This work was supported by the National Social Science Foundation of China (18ZDA309) and the National Natural Science Foundation of China (71531012). Thanks also go to the author's colleagues and students in preparing, handling and

organizing the data and documents, including Shimin Wang, Hua Shen, Jin Ruan, Shusen Zhang, Yang Xue, Xuan Zhang, Xiaoping Zhou, Yuefeng Ma, Jinshan Qi, Bo Wu, Mengdi Liu, etc.

Beijing, China Xun Liang

Contents

Part I Data

Part II AI Models and Algorithms

Part III Applications

Chapter 1
Introduction

1.1 Research Background

To date, people live life in online social network. We check our emails regularly and post messages on social media networks like Twitter, Facebook, etc. Our behaviors may bring about enormous data that hold the characteristics as 4Vs: volume, variety, velocity, and value. These data are of great importance for analyzing social group behavior and exploring potential huge social value. Therefore, in recent years, a new discipline emerged as the times require. The social computing is a field that leverages the ability to collect and analyze data at a scale that takes use of the potential online crowd wisdom. This is a data-driven interdisciplinary subject involving information science, behavioral science, psychology, systems engineering and decision science, which could provide various solutions to complex social problems [1].

With the outbreak of the big data online, social computing produced a variety of data acquisition modes, such as open data sharing in research communities, interactive data recording through video, people's links with email data. In addition, trajectory data, electronic communication data, online data, etc., which are all related to people's social behavior. Besides, social computing makes it possible to solve many traditional sociological problems efficiently. For example, complex problems can be greatly simplified by establishing system model and designing corresponding big data algorithm. According to the purpose of decision-making, a variety of schemes are designed and compared to provide the basis for the implementation of decision-making optimization. Beyond the space-time limit of the research object, the simulation of the research object is realized on the computer, and the tracking research or prediction research of the object which is difficult to be realized by other ways is completed. It is based on the study of objective factors and their interrelations, as well as the rigorous and accurate operation of computers. Therefore, it has considerable objectivity. In addition, a large number of advanced data analysis technologies emerged in the field of social computing.Big data analytics is the application of advanced analytics tools to big data sets. These analysis tools include prediction analysis, data mining, statistics, artificial intelligence, natural language processing,

X. Liang, *Social Computing with Artificial Intelligence*,
https://doi.org/10.1007/978-981-15-7760-4_1

machine learning, deep learning techniques, and so on. New analytical tools and technologies have enabled us to make better decisions. However, big data has also brought privacy issues subsequently. For example, in April 2018, Facebook CEO Zuckerberg went to a congressional hearing to receive information from senators about the suspected leak of 87 million user-sensitive information from Facebook to Cambridge Analytics to assist Trump in the election.

In any case, it is worth mentioning that in the field of social computing, researchers utilize massive network data to analyze the opinions of online user groups, excavating huge potential value and providing important help for decision-makers to make more effective decision analysis. This is quite difficult to achieve in traditional research. For example, Wang selected three well-known hot pot chain enterprises as the research object. According to the principle of MapReduce, he collected and processed the sample data, and adopted online analytical processing (OLAP) technology to achieve the visualization of customer network satisfaction in large data environment from three indicators: time, region, and satisfaction. Wang put forward that in the context of big data, banks can analyze and mine potential customers for existing credit card customers according to credit card transaction data, and upgrade credit cards at the same time. Through accurate database analysis, cross-selling can be carried out, new customers can be acquired, and lost users can be recovered. Mallinger and Stefl [2] and Lewis et al. [3] combined traditional methods with big data methods to study how people use data in decision-making, analyze user-generated content in the era of big data. They cannot only play the role of computing methods, cope with large data sets, but also ensure accuracy. Ceron et al. [4] obtained the emotional orientation data of netizens in the 2012 French elections through Twitter and predicted the results of the elections, which proved the better predictive ability of social media. Through collecting the public opinion information of “11.16 school bus accident” in Sina, Tencent and People’s Network, Kangwei constructed the network topology map of the incident, and then put forward the guiding strategy of the network public opinion. The Public Opinion Research Laboratory of Shanghai Jiaotong University innovated the research paradigm of public opinion, combined big data mining with social investigation, and constructed a “comprehensive public opinion research framework.” On the one hand, it utilized big data mining to find out the relationship between various factors in public opinion events and predicted the development trend of public opinion events. On the other hand, it explored individual’s cognition and evaluation of social hotspot issues through social investigation and transforms public opinion research from simplification, unilateralization, and staticization to panoramic, three-dimensional and dynamic.

In conclusion, we believe that confronted with massive online social media data, in order to fully tap the potential value of user groups, social computing future research areas will have the following trend characteristics: (1) Social computing research task needs to shift from the “causal relationship” of the data to the “relevant relationship.” In the social computing research activities, we no longer pay too much attention to “causality,” but put more emphasis on “correlation” [5].The most typical example is Google Flu Trends, which accurately estimates flu epidemics based on billions of search results around the world [6]. (2) Social computing research objects need to shift from single structured text data to multi-source, unstructured heterogeneous data. Traditional structured or semi-structured data has gradually turned to unstructured data composed of real-time text, audio, video, etc. Most of them are PB-level multi-source heterogeneous data streams, which are far beyond the scope of literature analysis, comparison, induction, and deduction, and are the contents of "not researching" or "difficult researching" in the past. (3) Social computing research methods need to shift from time-delay, static traditional analysis methods to real-time, dynamic, and interactive big data analysis methods. The traditional "micro-processing" method for small data is obviously incapable of facing the multi-type, multimedia, cross-time, cross-geographic, cross-language big data, and the user's individual needs. In the big data environment, the data analysis process emphasizes real-time, situational, and coordination. The platform for analysis requires higher scalability, fault tolerance, and support for heterogeneous data sources. (4) Social computing research process needs to shift from a "top-down" model to a "bottom-up" model. Traditional empirical research emphasizes the establishment of hypotheses, the collection of data, and the applicability of falsification theory under the premise of theory. It is a top-down analysis model. Random data sampling, data acquisition, verification hypothesis, questions that are not asked in the survey will not be answered, and the research on the target problem of graduate students will not be involved. In the big data environment, users are more urgently required to automatically identify valuable rules or patterns from massive data. It is a bottom-up mining model. (5) Social computing research applications need to be transferred from information decision services to knowledge decision services. Application services in a big data environment require addressing the problem of sparse value in big data, providing the most direct, reliable, real-time, and intelligent visualization solutions for the specific needs of business managers or government decision-makers, from traditional information services to knowledge service. Based on the above analysis, it is known that the innovation of social computing research method in the big data environment is the inevitable requirement for the development of network technology and network resources to a certain scale and degree. It has become an important topic of concern in the academic community and the industry, so it has a strong research significance and research value.

Therefore, this book came into being under this research trend. This book is aimed at social media network as the research object, taking online group analysis and value mining as the research goal, using machine learning-related algorithms to systematically sort out the data processing, method model and practical application of social computing research under big data environment. It should be pointed out that the purpose of this book is to provide a more systematic "data-model-application" theoretical research perspective for researchers in the field of social computing, so only the key areas and achievements of research are mentioned. We hope that readers can be inspired, and carry out further and in-depth research in a certain field under the guidance of a complete theoretical framework.

1.2 Mainstreams of Research Field

At present, the main research fields and methods of social computing focus on social data perception and knowledge discovery, individual and group community modeling, information dissemination and network dynamic evolution, decision support and application and other fields.

1.2.1 Social Data Perception and Knowledge Discovery

Social data acquisition and rule knowledge mining include social learning, social media analysis and information mining, emotion and opinion mining, behavior recognition and prediction. The main forms of social data include text, image, audio, video, etc. Its sources include not only network media information (including blogs, forums, news websites), but also private networks, traditional media, and closed source data of application departments. Knowledge discovery based on social data includes the analysis and mining of social individual or group behavior and psychology. Various learning algorithms have been used to predict organizational behavior. Based on behavioral prediction, the planning reasoning method can identify deep information such as the target and intention of the behavior. The psychological analysis of social groups is mainly oriented to text information (including texts transformed into speech recognition). By analyzing a large amount of social media information, it excavates the opinions and emotional tendencies of netizens.

1.2.2 Community Modeling of Individuals and Groups

Community modeling of individuals and groups includes the construction of behavior, cognitive and psychological models of social individuals or groups, and the analysis of behavior characteristics of social groups. It also includes the modeling of community structure, interaction patterns, and social relations among individuals. Many theoretical models of social sciences are related to the social modeling of individuals and groups. For example, social psychology reveals the formation mechanism of social cognition and psychology and its basic law of development. Social dynamics studies the dynamic process and evolution law of human social development. From the perspective of computating, the study of social individuals and groups is mostly based on textual data, and the trend of recent work is to analyze and model the characteristics of multimedia data and group behavior. Social network is the main means of describing the social interaction and interaction between individuals. The identification of social groups mainly finds potential social groups through the link relationship between network nodes.

1.2.3 Information Dissemination and Dynamic Evolution of Network

The research of information dissemination and network dynamic evolution is to analyze the characteristics of crowd interaction and the evolution of social events, including social network structure, information diffusion and impact, complex network and network dynamics, group interaction and collaboration. Computational sociology [7] holds that a large amount of information on the Internet, such as blogs, forums, chats, consumption records, e-mail, etc., are the mapping of human or organizational behavior in the real world in cyberspace. These network data can be used to analyze the behavior patterns of individuals and groups, thus deepening people's understanding of life, organization, and society. The study of computational sociology involves people's interactions, the form of social group networks, and their evolutionary laws. The evolutionary law analysis of social events mainly focuses on the analysis and evaluation of the process and mechanism of the occurrence, development, intensification, maintenance, and attenuation of social events. For example, in analyzing the evolution law of group activities, researchers use social dynamics to analyze the law of human space-time trajectory based on long-term tracking and detection of 100,000 mobile user terminals, and find that people's operation mode follows repeatable mode [8]. In addition, researchers have used a variety of models to analyze the dissemination, diffusion, and influencing factors of information in the network.

1.2.4 Decision Support and Application

The applications of social computing in the fields of social economy and security include providing decision support, emergency warning, policy evaluation and suggestions to managers and society. In recent years, social computing has made great progress and has been widely used. Network social media can often make a more rapid, sensitive, and accurate response than traditional media because it can fully reflect people's value orientation and true will. Open source information plays an important role in decision support and emergency warning. In the field of social and public security, the Intelligence and Informatics Research Team of the Institute of Automation of the Chinese Academy of Sciences cooperated with relevant national business departments to develop a large-scale open source intelligence acquisition and analysis processing system based on the ACP method, real-time monitoring, analysis and early warning of social intelligence. As well as decision support and services, it has been widely used in the actual business and security related fields of relevant departments. Socio-cultural computing has been applied to security and antiterrorism decision-making early warning [9, 10]. In addition, due to the complexity of social systems, large-scale social computing research needs computing environment and platform support, including cloud computing platform and various modeling, analysis, application, integration tools, and simulation environment.

1.3 Structure of This Book

This book integrates the four main research fields of social computing with artificial intelligence, namely supervised and unsupervised learning models, social data perception and knowledge discovery, individual and group community modeling, information dissemination and network dynamic evolution, decision support and application, into three research dimensions of data, methods, and applications, and forms the whole content of this book.

Furthermore, the book is divided into three parts: the first part, the data (including Chaps. 2 and 3), introduces the social computing data collection and data analysis algorithm, which roughly corresponds to the social data perception in the four fields of social computing. The second part, model (including Chaps. 4, 5, 6, 7 and 8), introduces supervised, unsupervised, and semi-supervised learning models, as well as more state-of-the-art artificial intelligence algorithms, the opinions content mining, community network structure, and social computing. The three aspects of social computing model of dynamic information dissemination roughly correspond to knowledge discovery, community modeling of individuals and groups, information dissemination, and network dynamic evolution in the four fields of social computing. The third part, application (including Chaps. 9, 10, 11 and 12), introduces the application cases of public security and emergency management, social computing application in business decision support, unsupervised oracle handwriting recognition, and social computing application in online crowd behavior and psychology.

Furthermore, for the specific content of each chapter is concerned, this chapter points out the research background, summarizes the research fields and main methods of social computing. Chapter 2 focuses on data-related content, including the source and classification of data in social computing, data acquisition methods and main tools, as well as data acquisition model and system platform. Chapter 3 introduces data processing principles and methods. Chapter 4 introduces the supervised and unsupervised machine learning method for data analysis. One of the major contributions of this book is to systematically summarize, sort out, and classify the current mainstream machine learning methods, and introduce them according to supervised learning, unsupervised learning and reinforcement learning. In addition, in recent years, deep learning has gained more outstanding performance in the field of social computing. Therefore, we list in-depth learning as a separate section to elaborate. Chapter 5 specifically introduces more state-of-the-art artificial intelligence algorithms, such as deep learning, reinforcement learning, brother learning, and epiphany learning. Chapter 6 introduces public opinion mining and analysis based on social network content mining. According to different content types, this chapter is further divided into text information mining analysis and network image mining analysis. Chapter 7 introduces the research of community discovery based on social network structure in social computing. It is further divided into the introduction of social network topology structure and model and the research of public opinion network community discovery. Chapter 8 introduces user role value mining based on network communication mechanism in social computing. This chapter introduces the analysis of the mechanism of network communication, the model of the influence of public opinion on users' behavior, and the research on the identification of users' roles in social networks. Chapter 9 introduces the application of social computing in emergency management decision support. Chapter 10 introduces the application of social computing in early risk warning in business decision support. Chapter 11 illustrates the unsupervised handwriting recognition method based on pic2vec and also discusses whether oracle bone inscriptions were engraved by right hand or left hand. At the end of the book, Chap. 12 presents social computing applications in online crowd behavior and psychology.

References

1. Lazer D, Pentland A, Adamic L (2009) Computational social science. Science 323(1):721–723
2. Mallinger M, Stefl M (2015) Big data decision making. Graziadio Bus Rev 18(2):25–33
3. Lewis SC, Zamith R, Hermida A (2013) Content analysis in an era of big data: a hybrid approach to computational and manual methods. J Broadcast Electron Media 57(1):34–52
4. Ceron A, Curini L, Iacus SM (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media Soc 16(2):340–358
5. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and thinking. Houghton Mifflin Harcourt, Boston, pp 50–72
6. Carneiro HA, Mylonakis E (2009) Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–1564
7. Lazer D, Pentland A, Adamic L (2009) Life in the network: the coming age of computational social science. Science 323(5915):721–723
8. González MC, Hidalgo CA, Barabási AL (2009) Understanding individual human mobility patterns. Nature 458(7235)
9. Subrahmanian VS (2007) Computer science. Cultural modeling in real time. Science 317(5844):1509
10. Subrahmanian VS, Albanese M, Martinez MV (2007) CARA: a cultural-reasoning architecture. IEEE Intell Syst 22(2):12–16

Part I
Data

Part I of the book is called *Data*, which includes Chaps. 2 and 3. As we all know, *Data* is a very important research foundation in all social computing studies. Therefore, we intentionally use two chapters to introduce data acquisition and data analysis algorithms in the field of social computing. Specifically, Chap. 2 briefly introduces data sources and classification, as well as data acquisition tools in the field of social computing. In Chap. 3, we address the data processing principles and methods.

Chapter 2
Data Collection

2.1 Data Types and Sources

In the aspect of data acquisition, traditional data acquisition methods, such as questionnaire survey, user interviews, experimental observation, etc., are usually used, which will result in shortcomings such as single data source, small magnitude, scarce type, and delayed information. Even in the Internet age, the method of online questionnaire survey will cause the problem of inaccurate information. In the big data environment, using the characteristics of large data, such as huge volume, variety and fast processing speed, will make up for the above shortcomings, so as to make the data sources of social computing more diverse, more magnitude, richer types, more time-consuming information, and provide better data support for social computing analysis and decision-making in the big data environment. Therefore, due to its extensive research fields and application scenarios, social computing has made its data sources an important feature of mass, diversity, heterogeneity, and timeliness. Generally speaking, social computing involves disparate data including text, images, speech, multimedia social media, and spatial, temporal, and spatiotemporal data.

The main sources of social computing data include communication companies, institutional databases, e-commerce platforms, social networks, search engines, and forums. Its data types include, but are not limited to, telephone communication records, enterprise and government management storage data, e-commerce website transaction data, social network platform and user data, search engine logs, web forum user information and comment data, etc. The data sources and classification of social computing are briefly summarized as shown in Table 2.1.

X. Liang, *Social Computing with Artificial Intelligence*,
https://doi.org/10.1007/978-981-15-7760-4_2

Table 2.1 Social computing data sources and classification

Institutional body	Openness	Type
Non-internet institutions	Private data	Transaction data of financial institutions, statistical data of government departments, telephone users' communication records
	Private data	Data published by government business organizations
Internet enterprises	Private data	E-commerce user transaction data, user registration information
	Private data	Social networking user reviews, product reviews, search engine logs, open data source websites

Private data sources are authoritative and confidential to the data, which is usually difficult to obtain. Relevant organizations often use the data processing and analysis department or cooperate with scientific research institutions to explore the value of data. Public data sources, such as Twitter, Facebook, and other social network information, Google search-related data, Amazon product user reviews, etc., often become the focus of various scholars because of their openness, relatively easy access, mass, heterogeneity.

2.2 Data Collection and Tools

2.2.1 Data Acquisition

For the acquisition of public data sources, there are generally three types. The first is to download the processed database directly from the established open database platform for research and application. We bring together several commonly used public databases as shown in Table 2.2. This list of public data sourcesis collected and tidyed from blogs, answers, and user responses. The table below is classified according to domains related to social computing, other extensive awesome sources can be found in this github repository[1]. Researchers can download databases directly from the website for machine learning training and testing (see Table 2.2).

The second is to crawl the information needed for research directly from the webpage. This type of data crawling process is often based on the Cascading Style Sheets (CSS) language to analyze the webpage and then capture the specific data required on the webpage. This type of data grabbing method should often be applied to extract specific data from web pages in the research process, but cannot find ready-made public data sources, and the website does not give the application programming interface website. For example, if a researcher wants to grab the content and comments posted by all users on Reddit from January to June 2018 on SubReddits

[1] https://github.com/awesomedata/awesome-public-datasets

called VACCINES (https://www.reddit.com/r/VACCINES/), then he would better crawl the content he needs by using a website crawl.

The third is data capture based on APIs of various websites. Many well-known websites provide dedicated data interfaces for developers and research. The most common websites include Twitter, Microformats, Mailboxes, LinkedIn, Google Buzz, Blogs, Facebook, etc. Researchers can find API documents and rules on relevant official websites. For example, Twitter gives API documents (http://apiwiki.twitter.com/). Through API interfaces, researchers can obtain many needed data. In addition, it should be noted that many researchers have further encapsulated the original API to form a more user-friendly toolkit, many of which are published in github. For example, a minimal wrapper around Twitter's webAPI is available through a package called Twitter (https://github.com/sixohsix/twitter).

2.2.2 Common Data Processing Toolkit

After introducing the sources of data classification and acquisition in the previous section, the next step is to consider whether to use different tools for data processing and analysis to achieve different research purposes. Based on Python programming language, we summarize the data processing toolkits commonly used in social computing data processing, which involve web crawler, text processing, scientific computing, machine learning and in-depth learning, as shown in Table 2.3.

Table 2.2 Commonly used public databases in the field of social computing

Category	Data set	Link	Description
Complex networks	Stanford large network data set collection	http://snap.stanford.edu/data/	SNAP
	Stanford longitudinal network data sources	http://stanford.edu/group/sonia/dataSources/index.html	SoNIA
	Stanford graph base	http://www3.cs.stonybrook.edu/~algorith/implement/graphbase/implement.shtml	–
	Koblenz network collection	http://konect.uni-koblenz.de/	KONECT
	NIST complex networks data collection	https://math.nist.gov/~RPozo/complex_datasets.html	NIST
	Scopus citation database	https://www.elsevier.com/solutions/scopus	SCOPUS
	UCI network data repository	https://networkdata.ics.uci.edu/resources.php	UCINET
	UFL sparse matrix collection	https://sparse.tamu.edu/	SuiteSparse
Data challenges	Kaggle competition data	https://www.kaggle.com/datasets	KAGGLE
	Netflix prize	https://netflixprize.com/leaderboard.html	NETFLIX
	Yelp data set challenge	https://www.yelp.com/dataset	YELP
	Telecom italia big data challenge	https://dandelion.eu/datamine/open-big-data/	Open big data

(continued)

Table 2.2 (continued)

Category	Data set	Link	Description
	Travistorrent data set—2017 mining challenge	https://travistorrent.testroots.org/	TravisTorrent
Machine learning	Keel repository for classification, regression and time series	http://sci2s.ugr.es/keel/datasets.php	Keel
	Machine learning data set repository	http://mldata.org/	MLDATA
	UCI machine learning repository	http://archive.ics.uci.edu/ml/index.php	UCIML
	Million song data set	https://labrosa.ee.columbia.edu/millionsong/	Million song
	Lending club loan data	https://www.lendingclub.com/info/download-data.action	Lending club
Natural language	WordNet	https://wordnet.princeton.edu/	WordNet
	DBpedia	http://wiki.dbpedia.org/Datasets	DBpedia
	Google books ngrams	https://aws.amazon.com/cn/datasets/google-books-ngrams/	Google books
	Gutenberg ebooks list	http://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs	Gutenberg
	Microsoft machine reading comprehension data set	http://www.msmarco.org/dataset.aspx	MS marco
	Machine translation of European languages	http://statmt.org/wmt11/translation-task.html#download	Machine translation

(continued)

Table 2.2 (continued)

Category	Data set	Link	Description
	Making sense of microposts 2016—named entity recognition	http://microposts2016.seas.upenn.edu/challenge.html	NEEL
	Multidomain sentiment data set (version 2.0)	http://www.cs.jhu.edu/~mdredze/datasets/sentiment/	Multidomain sentiment
	SMS spam collection in English	http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/	SMS spam
	Stanford question answering data set	https://rajpurkar.github.io/SQuAD-explorer/	SQuAD
Online website data sets	Amazon	http://aws.amazon.com/datasets	Amazon
	Reddit data sets	http://reddit.com/r/datasets	Reddit
	Archive.org datasets	https://archive.org/details/datasets	Archive
	Google	http://www.google.com/publicdata/directory	Google
	Data mining and data science	https://www.kdnuggets.com/datasets/index.html	KDNuggets
	Microsoft data science for research	https://www.microsoft.com/en-us/research/academic-program/data-science-microsoft-research/	Microsoft
	Yahoo webscope	https://webscope.sandbox.yahoo.com/	Yahoo
	Washington post list	http://www.washingtonpost.com/wp-srv/metro/data/datapost.html	Washington post
	Wikidata—wikipedia databases	https://www.wikidata.org/wiki/Wikidata:Database_download	WikiData

Table 2.3 Python data processing toolkit related to social computing

Category	Package	Link	Description
Webcrawler toolkit	Scrapy	http://scrapy.org/	Afast high-level screen scraping and web crawling framework for Python
	Beautiful Soup	http://www.crummy.com/software/BeautifulSoup/	Beautifu Soup is not a complete set of crawler tools, it needs to be used with urllib, but a set of HTML/XML data analysis, cleaning and acquisition tools
	Python-Goose	https://github.com/grangier/python-goose	Html content/article extractor, webscrapping lib in Python
Machine learning–data mining	Scikit-learn	http://scikit-learn.org/	It features various classification, regression and clustering algorithms
	Pandas	http://pandas.pydata.org/	Pandas is a software library written for the Python programming language for data manipulation and analysis
	Mlpy	http://mlpy.sourceforge.net/	Mlpy is a Python module for machine learning built on top of NumPy/SciPy and the GNU Scientific Libraries
	MDP	http://mdp-toolkit.sourceforge.net/	Modular toolkit for data processing (MDP) is a Python data processing framework

(continued)

Table 2.3 (continued)

	PyBrain	http://www.pybrain.org/	Its goal is to offer flexible, easy-to-use yet still powerful algorithms for machine learning tasks and a variety of predefined environments to test and compare your algorithms
	NetworkX	https://networkx.github.io/	NetworkX is a graph theory and complex network modeling tool developed in Python language. It has built-in common graph and complex network analysis algorithm, which can easily carry out complex network data analysis, simulation modeling and so on
Deep learning	TensorFlow	https://www.tensorflow.org	TensorFlow is a system that transfers complex data structures to artificial intelligence neural networks for analysis and processing. It can be used in many fields of machine learning and deep learning, such as speech recognition or image recognition
	Caffe	http://caffe.berkeleyvision.org	Convolutional architecture for fast feature embedding is a commonly used deep learning framework, which is widely used in video and image processing

(continued)

Table 2.3 (continued)

	Theano	http://www.deeplearning.net/software/theano/	Theano is a Python library that allows youdefine, optimize, and evaluate mathematical expressions involving multidimensional arrays efficiently
	PyTorch	http://pytorch.org/	PyTorch is a deep learning tensor library optimized with GPU and CPU
	Pylearn2	http://deeplearning.net/software/pylearn2	Pylearn2 is based on Theano and partly depends on scikit-learning. Pylearn2 will be able to process vector, image, video and other data, and provide in-depth learning models such as MLP, RBM, and SDA

Chapter 3
Data Processing Methodology

3.1 Data Processing Principles

In the existing general big data, structured data only account for about 15%, and the remaining 85% are unstructured data. Under the research of social computing, the proportion of unstructured data will be higher, so it is more necessary to reconsider and design corresponding methods to solve the specific problems encountered in the social computing big data environment. On the one hand, it is necessary to study and discuss through multidisciplinary intersections including mathematics, economics, sociology, computer science, and management science, and to define the individual manifestations, the general characteristics and the basic principles of unstructured and semi-structured data in social computing. On the other hand, each representation of big data only presents the side performance of the data itself, not the whole picture. How to use appropriate data representation and acquisition form for the specific research methods of social computing, and propose a data fusion method for social computing research, will also be an important part of the research. Therefore, after explaining the data classification and source, data acquisition and tools, we will briefly introduce some unique ideas and methods of data acquisition related to social computing in this section.

3.1.1 Behavior Tracking

In the era of big data, people's various behavioral data can be tracked through web logs. The tracking of network user behavior logs is fundamentally different from data acquisition in traditional research. In the past research, researchers mostly used passive questionnaires or surveys to obtain users' data information. This acquisition mode not only makes the amount of data acquired smaller, fewer types, but also makes the reliability of data acquisition relatively low. In the era of big data, we can ensure the timeliness of data to the greatest extent by tracking the historical data of

X. Liang, *Social Computing with Artificial Intelligence*,
https://doi.org/10.1007/978-981-15-7760-4_3