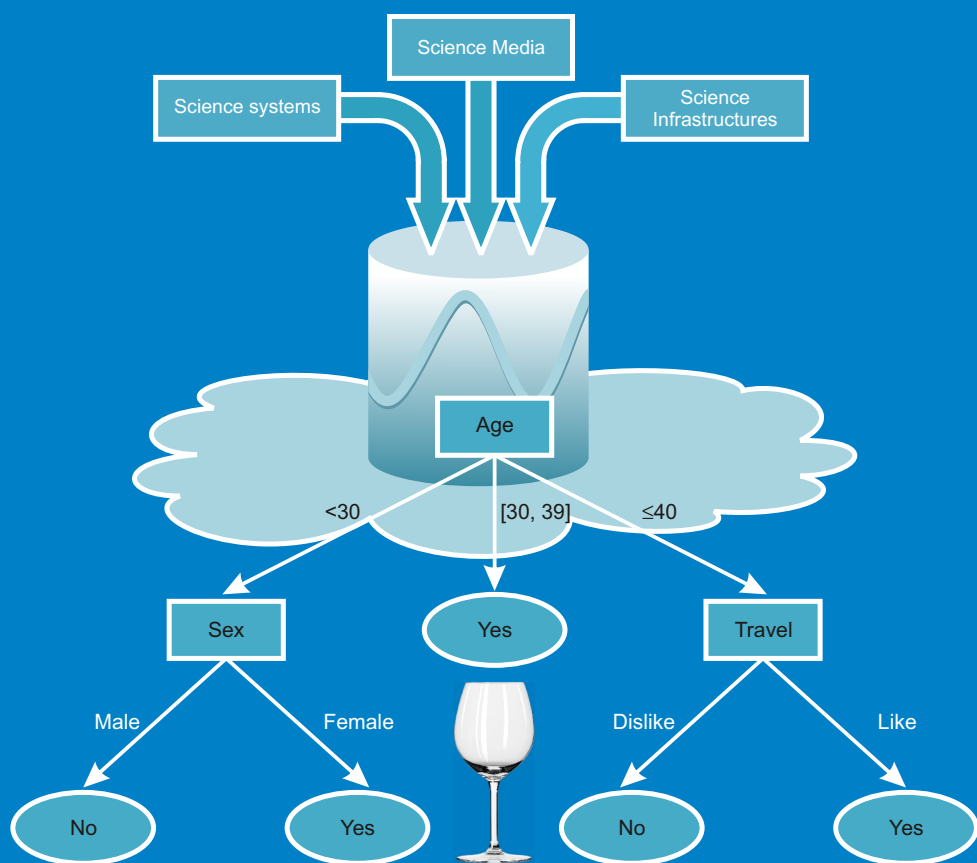


Social Big Data Mining

Hiroshi Ishikawa



CRC Press
Taylor & Francis Group

A SCIENCE PUBLISHERS BOOK

Social Big Data Mining

Social Big Data Mining

Hiroshi Ishikawa

Dr. Sci., Prof.

Information and Communication Systems

Faculty of System Design

Tokyo Metropolitan University

Tokyo, Japan



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A SCIENCE PUBLISHERS BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2015 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20150218

International Standard Book Number-13: 978-1-4987-1094-7 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Preface

In the present age, large amounts of data are produced continuously in science, on the internet, and in physical systems. Such data are collectively called data deluge. According to researches carried out by IDC, the size of data which are generated and reproduced all over the world every year is estimated to be 161 exa bytes. The total amount of data produced in 2011 exceeded 10 or more times the storage capacity of the storage media available in that year.

Experts in scientific and engineering fields produce a large amount of data by observing and analyzing the target phenomena. Even ordinary people voluntarily post a vast amount of data via various social media on the internet. Furthermore, people unconsciously produce data via various actions detected by physical systems in the real world. It is expected that such data can generate various values.

In the above-mentioned research report of IDC, data produced in science, the internet, and in physical systems are collectively called big data.

The features of big data can be summarized as follows:

- The quantity (Volume) of data is extraordinary, as the name denotes.
- The kinds (Variety) of data have expanded into unstructured texts, semi-structured data such as XML, and graphs (i.e., networks).
- As is often the case with Twitter and sensor data streams, the speed (Velocity) at which data are generated is very high.

Therefore, big data is often characterized as V^3 by taking the initial letters of these three terms Volume, Variety, and Velocity. Big data are expected to create not only knowledge in science but also derive values in various commercial ventures.

“Variety” implies that big data appear in a wide variety of applications. Big data inherently contain “vagueness” such as inconsistency and deficiency. Such vagueness must be resolved in order to obtain quality analysis results. Moreover, a recent survey done in Japan has made it clear that a lot of users have “vague” concerns as to the securities and mechanisms of big data applications. The resolution of such concerns is one of the keys

to successful diffusion of big data applications. In this sense, V^4 should be used to characterise big data, instead of V^3 .

Data analysts are also called data scientists. In the era of big data, data scientists are more and more in demand. The capabilities and expertise necessary for big data scientists include:

- Ability to construct a hypothesis
- Ability to verify a hypothesis
- Ability to mine social data as well as generic Web data
- Ability to process natural language information
- Ability to represent data and knowledge appropriately
- Ability to visualize data and results appropriately
- Ability to use GIS (geographical information systems)
- Knowledge about a wide variety of applications
- Knowledge about scalability
- Knowledge and follow ethics and laws about privacy and security
- Can use security systems
- Can communicate with customers

This book is not necessarily comprehensive according to the above criteria. Instead, from the viewpoint of social big data, this book focusses on the basic concepts and the related technologies as follows:

- Big data and social data
- The concept of a hypothesis
- Data mining for making a hypothesis
- Multivariate analysis for verifying the hypothesis
- Web mining and media mining
- Natural language processing
- Social big data applications
- Scalability

In short, featuring hypotheses, which are supposed to have an ever-increasingly important role in the era of social big data, this book explains the analytical techniques such as modeling, data mining, and multivariate analysis for social big data. It is different from other similar books in that it aims to present the overall picture of social big data from fundamental concepts to applications while standing on academic bases.

I hope that this book will be widely used by readers who are interested in social big data, including students, engineers, scientists, and other professionals. In addition, I would like to deeply thank my wife Tazuko, my children Takashi and Hitomi for their affectionate support.

July, 2014

Hiroshi Ishikawa
Kakio, Dijon and Bayonne

Contents

<i>Preface</i>	v
1. Social Media	1
2. Big Data and Social Data	16
3. Hypotheses in the Era of Big Data	46
4. Social Big Data Applications	66
5. Basic Concepts in Data Mining	86
6. Association Rule Mining	99
7. Clustering	111
8. Classification	125
9. Prediction	136
10. Web Structure Mining	149
11. Web Content Mining	165
12. Web Access Log Mining, Information Extraction, and Deep Web Mining	185
13. Media Mining	201
14. Scalability and Outlier Detection	228
<i>Appendix I: Capabilities and Expertise Required for Data Scientists in the Age of Big Data</i>	243
<i>Appendix II: Remarks on Relationships Among Structure-, Content-, and Access Log Mining Techniques</i>	247
<i>Index</i>	249
<i>Color Plate Section</i>	255

1

Social Media

Social media are indispensable elements of social big data applications. In this chapter, we will first classify social media into several categories and explain the features of each category in order to better understand what social media are. Then we will select important media categories from a viewpoint of analysis required for social big data applications, address representative social media included in each category, and describe the characteristics of the social media, focusing on the statistics, structures, and interactions of social media as well as the relationships with other similar social media.

1.1 What are Social Media?

Generally, a social media site consists of an information system as its platform and its users on the Web. The system enables the user to perform direct interactions with it. The user is identified by the system along with other users as well. Two or more users constitute explicit or implicit communities, that is, social networks. The user in social media is generally called an actor in the context of social network analysis. By participating in the social network as well as directly interacting with the system, the user can enjoy services provided by the social media site.

More specifically, social media can be classified into the following categories based on the service contents.

- *Blogging*: Services in this category enable the user to publish explanations, sentiments, evaluations, actions, and ideas about certain topics including personal or social events in a text in the style of a diary.
- *Micro blogging*: The user describes a certain topic frequently in shorter texts in micro blogging. For example, a tweet, an article of Twitter, consists of at most 140 characters.

- *SNS (Social Network Service)*: Services in this category literally support creating social networks among users.
- *Sharing service*: Services in this category enable the user to share movies, audios, photographs, and bookmarks.
- *Video communication*: The users can hold a meeting and chat with other users using live videos as services in this category.
- *Social search*: Services in this category enable the user to reflect the likings and opinions of current search results in the subsequent searches. Other services allow not only experts but also users to directly reply to queries.
- *Social news*: Through services in this category the user can contribute news as a primary source and can also re-post and evaluate favorite news items which have already been posted.
- *Social gaming*: Services in this category enable the user to play games with other users connected by SNS.
- *Crowd sourcing*: Through services in this category, the user can outsource a part or all of his work to outside users who are capable of doing the work.
- *Collaboration*: Services in this category support cooperative work among users and they enable the users to publish a result of the cooperative work.

1.2 Representative Social Media

In consideration of user volumes and the social impact of media in the present circumstances, micro blogging, SNS, movie sharing, photograph sharing, and collaboration are important categories of social big data applications, where social media data are analyzed and the results are utilized as one of big data sources. The profiles (i.e., features) of representative social media in each category will be explained as well as generic Web, paying attention to the following aspects which are effective for analysis:

- Category and foundation
- Numbers
- Data structures
- Main interactions
- Comparison with similar media
- API

1.2.1 Twitter

(1) Category and foundation

Twitter [Twitter 2014] [Twitter-Wikipedia 2014] is one of the platform services for micro blogging founded by Jack Dorsey in 2005 (see Fig. 1.1).

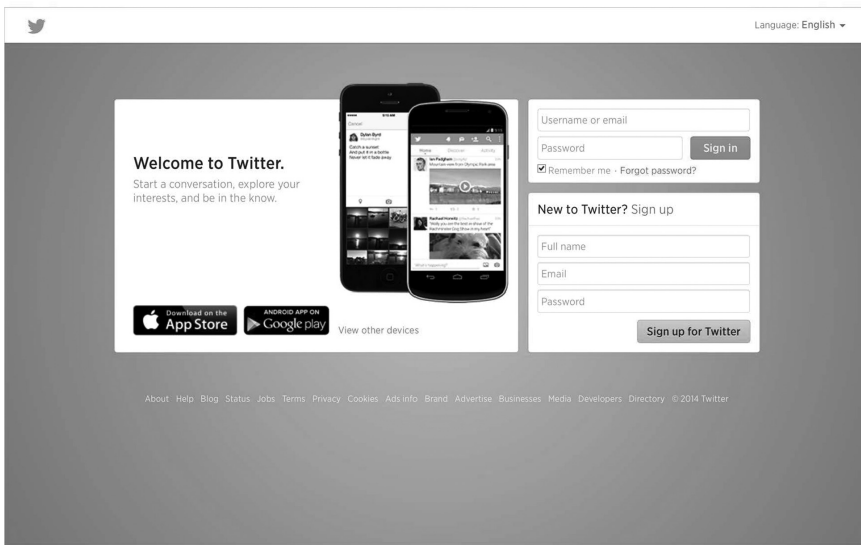


Figure 1.1 Twitter.

Color image of this figure appears in the color plate section at the end of the book.

Twitter started from the ideas about development of media which are highly live and suitable for communication among friends. It is said that it has attracted attention partly because its users have increased so rapidly. For example, in Japan, when the animation movie “Castle in the Sky” by Hayao Miyazaki was broadcast as a TV program in 2011, there were 25,088 tweets in one second, which made it the center of attention.

(2) Numbers

- Active users: 200 M (M: Million)
- The number of searches per day: 1.6 B (B: Billion)
- The number of tweets per day: 400 M

(3) Data structures

(Related to users)

- Account
- Profile

(Related to contents)

- Tweet

(Related to relationships)

- Links to Web sites, video, and photo
- The follower-followee relationship between users

- Memory of searches
- List of users
- Bookmark of tweets

(4) Main interactions

- Creation and deletion of an account.
- Creation and change of a profile.
- Contribution of a tweet: Tweets contributed by a user who are followed by another user appear in the time line of the follower.
- Deletion of a tweet.
- Search of tweets: Tweets can be searched with search terms or user names.
- Retweet: If a tweet is retweeted by a user, the tweet will appear in the time line of the follower. In other words, if the user follows another user and the latter user retweets a certain tweet, then the tweet will appear in the time line of the former user.
- Reply: If a user replies to a message by user who contributed the tweet, then the message will appear in the time line of another user who follows both of them.
- Sending a direct message: The user directly sends a message to its follower.
- Addition of location information to tweets.
- Inclusion of hash tags in a tweet: Tweets are searched with the character string starting with “#” as one of search terms. Hash tags often indicate certain topics or constitute coherent communities.
- Embedding URL of a Web page in a tweet.
- Embedding of a video as a link to it in a tweet.
- Upload and sharing of a photo.

(5) Comparison with similar media

Twitter is text-oriented like general blogging platforms such as WordPress [WordPress 2014] and Blogger [Blogger 2014]. Of course, tweets can also include links to other media as described above. On the other hand, the number of characters of tweets is less than that of general blog articles and tweets are more frequently posted. Incidentally, WordPress is not only a platform of blogging, but it also enables easy construction of applications upon LAMP (Linux Apache MySQL PHP) stacks, therefore it is widely used as CMS (Content Management System) for enterprises.

(6) API

Twitter offers REST (Representational State Transfer) and streaming as its Web services API.

1.2.2 Flickr

(1) Category and foundation

Flickr [Flickr 2014] [Flickr–Wikipedia 2014] is a photo sharing service launched by Ludicorp, a company founded by Stewart Butterfield and Caterina Fake in 2004 (see Fig. 1.2). Flickr focused on a chat service with

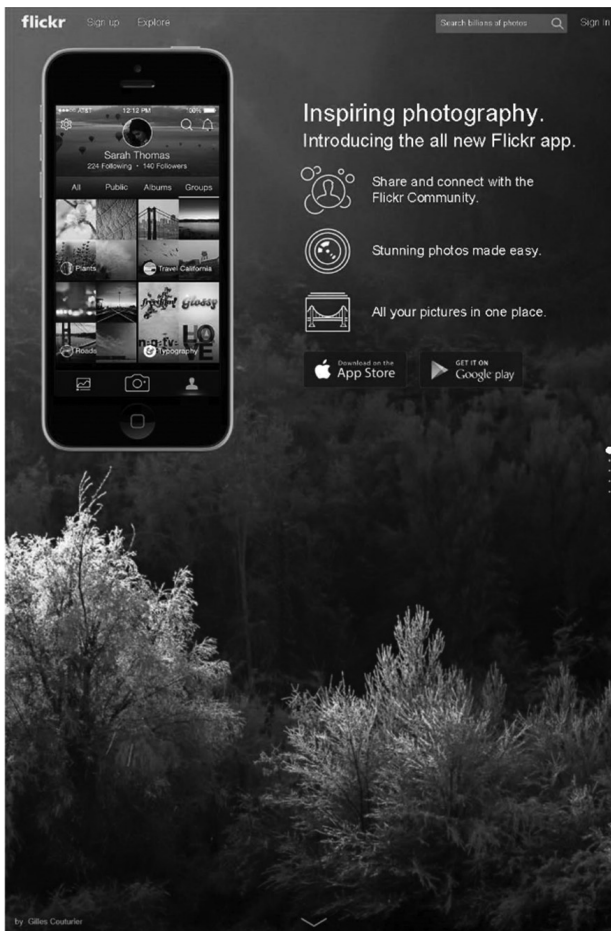


Figure 1.2 Flickr.

Color image of this figure appears in the color plate section at the end of the book.

real-time photo exchange in its early stages. However, the photo sharing service became more popular and the chat service, which was originally the main purpose, disappeared, partly because it had some problems.

(2) Numbers

- Registered users: 87 M
- The number of photos: 6 B

(3) Data structures

(Related to user)

- Account
- Profile

(Related to contents)

- Photo
- Set collection of photos
- Favorite photo
- Note
- Tag
- Exif (Exchangeable image file format)

(Related to relationships)

- Group
- Contact
- Bookmark of an album (a photo)

(4) Main interactions

- Creation and deletion of an account.
- Creation and change of a profile.
- Upload of a photo.
- Packing photos into a set collection.
- Appending notes to a photo.
- Arranging a photo on a map.
- Addition of a photo to a group.
- Making relationships between friends or families from contact.
- Search by explanation and tag.

(5) Comparisons with similar media

Although Picasa [Picasa 2014] and Photobucket [Photobucket 2014] are also popular like Flickr in the category of photo sharing services, here we will

take up Pinterest [Pinterest 2014] and Instagram [Instagram 2014] as new players which have unique features. Pinterest provides lightweight services on the user side compared with Flickr. That is, in Pinterest, the users can not only upload original photos like Flickr, but can also stick their favorite photos on their own bulletin boards by pins, which they have searched and found on Pinterest as well as on the Web. On the other hand, Instagram offers the users many filters by which they can edit photos easily. In June, 2012, an announcement was made that Facebook acquired Instagram.

(6) API

Flickr offers REST, XML-RPC (XML-Remote Procedure Call), and SOAP (originally, Simple Object Access Protocol) as Web service API.

1.2.3 YouTube

(1) Category and foundation

YouTube [YouTube 2014] [YouTube-Wikipedia 2014] is a video sharing service founded by Chad Hurley, Steve Chen, Jawed Karim, and others in 2005 (see Fig. 1.3). When they found difficulties in sharing videos which had recorded a dinner party, they came up with the idea of YouTube as a simple solution.

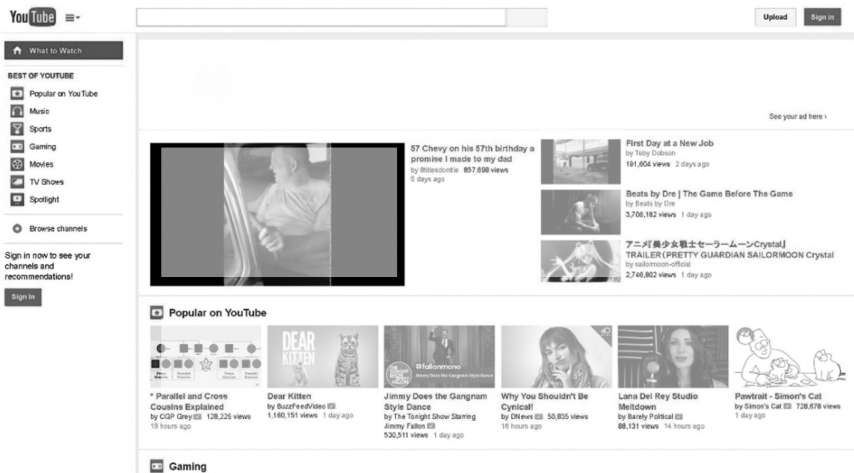


Figure 1.3 YouTube.

Color image of this figure appears in the color plate section at the end of the book.

(2) Numbers

- 100 hours of movies are uploaded every minute.
- More than 6 billion hours of movies are played per month.
- More than 1 billion users access per month.

(3) Data structures

(Related to users)

- Account

(Related to contents)

- Video
- Favorite

(Related to relationships)

- Channel

(4) Main interactions

- Creation and deletion of an account
- Creation and change of a profile
- Uploading a video
- Editing a video
- Attachment of a note to a video
- Playing a video
- Searching and browsing a video
- Star-rating of a video
- Addition of a comment to a video
- Registration of a channel in a list
- Addition of a video to favorite
- Sharing of a video through e-mail and SNS

(5) Comparison with similar media

As characteristic rivals, Japan-based Niconico (meaning smile in Japanese) [Niconico 2014] and the US-based USTREAM [USTREAM 2014] are picked up in this category. Although the Niconico Douga, one of the services provided by Niconico, is similar to YouTube, Niconico Douga allows the user to add comments to movies which can be superimposed on the movies and seen by other users later, unlike YouTube. Such comments in Niconico Douga have attracted a lot of users as well as the original contents. Niconico Live is another service provided by Niconico and is similar to the live video service of USTREAM. USTREAM was originally devised as a way by which US soldiers serving in the war with Iraq could communicate with their families. The function for posting tweets simultaneously with video

viewing made USTREAM popular. Both USTREAM and Niconico Live can be viewed as a new generation of broadcast services which are more targeted than the conventional mainstream services.

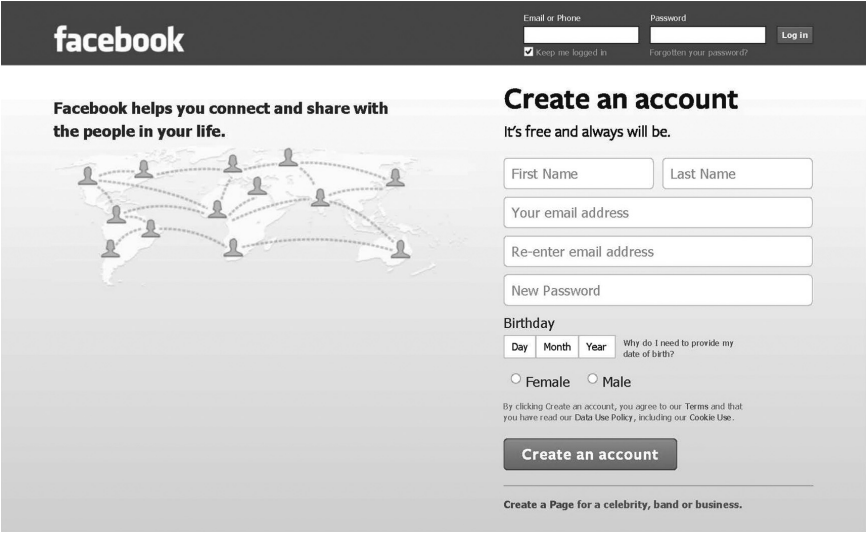
(6) API

YouTube provides the users with a library which enables the users to invoke its Web services from programming environments, such as Java and PHP.

1.2.4 Facebook

(1) Category and foundation

Facebook [Facebook 2014] [Facebook–Wikipedia 2014] is an integrated social networking service founded by Mark Zuckerberg and others in 2004, where the users participate in social networking under their real names (see Fig. 1.4). Facebook began from a site which was intended to promote exchange among students and has since grown to be a site which may affect fates of countries. Facebook has successfully promoted development of applications for Facebook by opening wide its development platform to application developers or giving them subsidies. Furthermore, Facebook has invented a mechanism called social advertisements. By Facebook's social advertisements, for example, the recommendation "your friend F



facebook

Email or Phone Password **Log In**

☒ Keep me logged in [Forgotten your password?](#)

Facebook helps you connect and share with the people in your life.

Create an account

It's free and always will be.

First Name Last Name

Your email address

Re-enter email address

New Password

Birthday

Day Month Year Why do I need to provide my date of birth?

☐ Female ☐ Male

By clicking Create an account, you agree to our Terms and that you have read our Data Use Policy, including our Cookie Use.

Create an account

Create a Page for a celebrity, band or business.

Figure 1.4 Facebook.

Color image of this figure appears in the color plate section at the end of the book.

purchased the product P'' will appear on the page of the user who is a friend of F. Facebook's social advertisements are distinguished from anonymous recommendations based on historical mining of customers behavior such as that of Amazon.

(2) Numbers

- Active users: 1 B

(3) Data structures

(Related to user)

- Account
- Profile

(Related to contents)

- Photo
- Video

(Related to relationships)

- Friend list
- News feed
- Group

(4) Main interactions

- Creation and deletion of an account
- Creation and update of a profile
- Friend search
- Division of friends into lists
- Connection search
- Contribution (recent status, photo, video, question)
- Display of time line
- Sending and receiving of a message

(5) Comparison with similar media

In addition to the facilities to include photos and videos like Flickr or YouTube, Facebook has also adopted the timeline function which is a basic facility of Twitter. Facebook incorporates the best of social media in other categories, so to say, a more advanced hybrid SNS as a whole.

(6) API

FQL (Facebook Query Language) is provided as API for accessing open graphs (that is, social graphs).

1.2.5 Wikipedia

(1) Category and foundation

Wikipedia [Wikipedia 2014] is an online encyclopedia service which is a result of collaborative work, founded by Jimmy Wales and Larry Sanger in 2001 (see Fig. 1.5). The history of Wikipedia began from Nupedia [Nupedia 2014], a project prior to it in 2000. Nupedia aimed at a similar online encyclopedia based on copyright-free contents. Unlike Wikipedia, however, Nupedia had adopted the traditional editorial processes for publishing articles based on the contributions and peer reviews by specialists. Originally, Wikipedia was constructed by a Wiki software for the purpose of increasing articles as well as contributors for Nupedia in 2001. In the early stages of Wikipedia, it earned its reputation through electric word-of-mouth and attracted a lot of attention through being mentioned in Slashdot



Figure 1.5 Wikipedia.

[Slashdot 2014], a social news site. Wikipedia has rapidly expanded its visitor attraction with the aid of search engines such as Google.

(2) Numbers

- Number of articles: 4 M (English-language edition)
- Number of users: 20 + M (English-language edition)

(3) Data structures

(Related to users)

- Account

(Related to contents)

- Page

(Related to relationship)

- Link

(4) Main interactions

(Administrator or editor)

- Creation, update, and deletion of an article
- Creation, update, and deletion of link to an article
- Change management (a revision history, difference)
- Search
- User management

(General user)

- Browse Pages in the site
- Search Pages in the site

(5) Comparison with similar media

From a viewpoint of platforms for collaboration, Wikipedia probably should be compared with other wiki media or cloud services (e.g., ZOHO [ZOHO 2014]). However, from another viewpoint of “search of knowledge” as the ultimate purpose of Wikipedia, players for social search services will be rivals of Wikipedia. You should note that differences between major search engines (e.g., Google [Google 2014] and Bing [Bing 2014]) and Wikipedia is being narrowed. Conventionally, such conventional search engines mechanically rank the search results and display them to the users. However, by allowing the users to intervene between search processes in certain forms, the current search engines are going to improve the quality of search results. Some search engines include relevant pages linked by friends in social media in search results. In order to get answers to a query, other

search engines discover people likely to answer the query from friends in social media or specialists on the Web, based on their profiles, uploaded photos, and blog articles.

(6) API

In Wikipedia, REST API of MediaWiki [MediaWiki 2014] can be used for accessing the Web services.

1.2.6 Generic Web

(1) Category and foundation

When Tim Berners-Lee joined CERN as a fellow, he came up with the prototype of the Web as a mechanism for global information sharing and created the first Web page in 1990. The next year, the outline of the WWW project was released and its service was started. Since the Web, in a certain sense, is the entire world in which we are interested, it contains all the categories of social media.

(2) Numbers

- The size of the indexable Web: $11.5 + B$ [Gulli et al. 2005]

(3) Data Structures

(Related to users)

NA

(Related to contents)

- Page

(Related to relationships)

- Link

(4) Main operations

(Administrator)

- Creation, update, and deletion of a page
- Creation, update, and deletion of a link

(General user)

- Page browse in a site
- Page search in a site
- Form input

(5) Comparisons with similar media

Since the Web is a universal set containing all the categories, we cannot compare it with other categories. Generally, the Web can be classified into the surface Web and the deep Web. While the sites of the surface Web allow the user to basically follow links and scan pages, those of the deep Web with back-end databases, create pages dynamically and display them to the user, based on the result of the database query which the user issues through the search form. Moreover, the sites of the deep Web are increasing rapidly [He et al. 2007]. The categories in the deep Web include on-line shopping services represented by Amazon, and various kinds of social media described in this book.

(6) API

Web services API provided by search engines such as Yahoo! can facilitate search of Web pages. Unless we use such API, we need to carry tedious Web crawling by ourselves.

1.2.7 Other social media

The categories of social media which have not yet been discussed will be enumerated below.

- **Sharing service:** In addition to photos and videos described previously, audios (e.g., Rhapsody [Rhapsody 2014], iTunes [iTunes 2014]) and bookmarks (e.g., Delicious [Delicious 2014], Japan-based Hatena bookmark [Hatena 2014]) are shared by users.
- **Video communication:** Users can communicate with each other through live videos. Skype [Skype 2014] and Tango [Tango 2014] are included in this category.
- **Social news:** The users can post original news or repost existing news by adding comments to them. Representative media of this category include Digg [Digg 2014] and Reddit [Reddit 2014] in addition to Slashdot.
- **Social gaming:** A group of users can play online games. The services in this category include FarmVille [FarmVille 2014] and Mafia Wars [Mafia Wars 2014].
- **Crowd sourcing:** The services in this category allow personal or enterprise users to outsource the whole or parts of a job to crowds in online communities. Amazon Mechanical Turk [Amazon Mechanical Turk 2014] for requesting labor-oriented work and InnoCentive [InnoCentive 2014] for requesting R&D-oriented work are included by the services in this category.

References

- [Amazon Mechanical Turk 2014] Amazon Mechanical Turk: Artificial Intelligence <https://www.mturk.com/mturk/welcome> accessed 2014
- [Bing 2014] Bing <http://www.bing.com> accessed 2014
- [Blogger 2014] Blogger <https://www.blogger.com> accessed 2014
- [Delicious 2014] Delicious <http://delicious.com> accessed 2014
- [Digg 2014] Digg <http://digg.com> accessed 2014
- [Facebook 2014] Facebook <https://www.facebook.com/> accessed 2014
- [Facebook–Wikipedia 2014] Facebook–Wikipedia <http://en.wikipedia.org/wiki/Facebook> accessed 2014
- [FarmVille 2014] FarmVille <http://company.zynga.com/games/farmville> accessed 2014
- [Flickr 2014] Flickr <https://www.flickr.com> accessed 2014
- [Flickr–Wikipedia 2014] Flickr–Wikipedia <http://en.wikipedia.org/wiki/Flickr> accessed 2014
- [Google 2014] Google <https://www.google.com> accessed 2014
- [Gulli et al. 2005] A. Gulli and A. Signorini: The indexable web is more than 11.5 billion pages. In Special interest tracks and posters of the 14th international conference on World Wide Web (WWW '05). ACM 902–903 (2005).
- [Hatena 2014] Hatena <http://www.hatena.ne.jp/> accessed 2014
- [He et al. 2007] Bin He, Mitesh Patel, Zhen Zhang and Kevin Chen-Chuan Chang: Accessing the deep web, Communications of the ACM 50(5): 94–101 (2007).
- [InnoCentive 2014] InnoCentive <http://www.innocentive.com> accessed 2014
- [Instagram 2014] Instagram <http://instagram.com/> accessed 2014
- [iTunes 2014] iTunes <https://www.apple.com/itunes/> accessed 2014
- [Mafia Wars 2014] Mafia Wars <http://www.mafiawars.com/> accessed 2014
- [MediaWiki 2014] MediaWiki <http://www.mediawiki.org/wiki/MediaWiki> accessed 2014
- [Niconico 2014] Niconico <http://www.nicovideo.jp/?header> accessed 2014
- [Nupedia 2014] Nupedia <http://en.wikipedia.org/wiki/Nupedia> accessed 2014
- [Picasa 2014] Picasa <https://www.picasa.google.com> accessed 2014
- [Photobucket 2014] Photobucket <http://photobucket.com/> accessed 2014
- [Pinterest 2014] Pinterest <https://www.pinterest.com/> accessed 2014
- [Reddit 2014] Reddit <http://www.reddit.com> accessed 2014
- [Rhapsody 2014] Rhapsody <http://try.rhapsody.com/> accessed 2014
- [Skype 2014] Skype <http://skype.com> accessed 2014
- [Slashdot 2014] Slashdot <http://www.slashdot.org> accessed 2014
- [Tango 2014] Tango <http://www.tango.me> accessed 2014
- [Twitter 2014] Twitter <https://twitter.com> accessed 2014
- [Twitter–Wikipedia 2014] Twitter–Wikipedia <http://en.wikipedia.org/wiki/Twitter> accessed 2014
- [USTREAM 2014] USTREAM <http://www.ustrea.tv> accessed 2014
- [Wikipedia 2014] Wikipedia <https://wikipedia.org> accessed 2014
- [WordPress 2014] WordPress <https://wordpress.com> accessed 2014
- [YouTube 2014] YouTube <http://www.youtube.com> accessed 2014
- [YouTube–Wikipedia 2014] YouTube–Wikipedia <http://en.wikipedia.org/wiki/YouTube> accessed 2014
- [ZOH0 2014] ZOH0 <https://www.zoho.com/> accessed 2014

2

Big Data and Social Data

At this moment, data deluge is continuously producing a large amount of data in various sectors of modern society. Such data are called big data. Big data contain data originating both in our physical real world and in social media. If both kinds of data are analyzed in a mutually related fashion, values which cannot be acquired only by independent analysis will be discovered and utilized in various applications ranging from business to science. In this chapter, modeling and analyzing interactions involving both the physical real world and social media as well as the technology enabling them will be explained. Data mining required for analysis will be explained in Part II.

2.1 Big Data

In the present age, large amounts of data are produced every moment in various fields, such as science, Internet, and physical systems. Such phenomena collectively called data deluge [Mcfedries 2011]. According to researches carried out by IDC [IDC 2008, IDC 2012], the size of data which are generated and reproduced all over the world every year is estimated to be 161 exa bytes (see Fig. 2.1). Here, kilo, mega, giga, tera, peta, exa, zetta are metric prefixes that increase by a factor of 10^3 . Exa and Zetta are the 18th power of 10 and the 21st power of 10, respectively. It is predicted that the total amount of data produced in 2011 exceeded 10 or more times the storage capacity of the storage media available in that year.

Astronomy, environmental science, particle physics, life science, and medical science are among the fields of science which produce a large amount of data by observation and analysis of the target phenomena. Radio telescopes, artificial satellites, particle accelerators, DNA sequencers, and MRIs continuously provide scientists with a tremendous amount of data.

Nowadays, even ordinary people, not to mention experts, produce a large amount of data directly and intentionally through the Internet

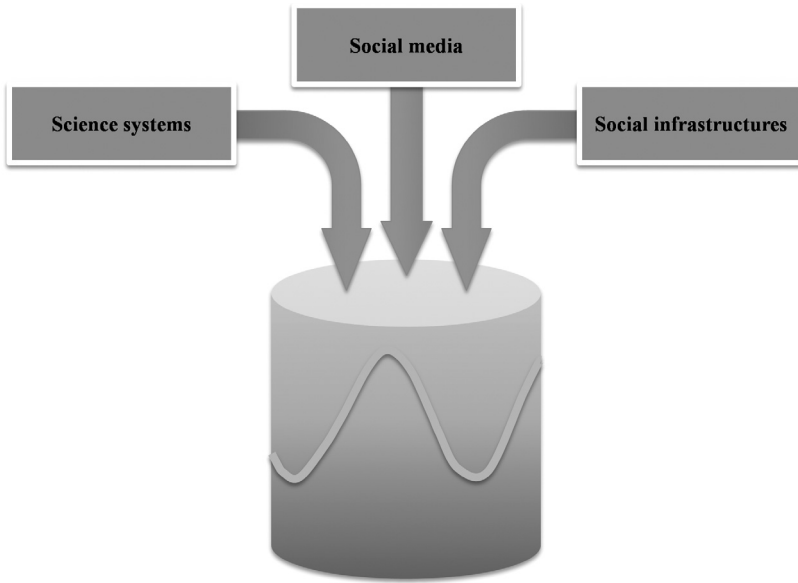


Figure 2.1 Data deluge.

services. Digital libraries, news, Web, Wiki, and social media. Twitter, Flickr, Facebook, and YouTube are representatives of the social media which have evolved rapidly in recent years. Moreover, some news sites such as Slashdot and some Wikis such as Wikipedia can be viewed as kinds of social media.

On the other hand, data originating in electric power apparatus, gas apparatus, digital cameras, surveillance video cameras, and sensors within buildings (e.g., passive infrared-, temperature-, illuminance-, humidity-, and carbon dioxide-sensors) and data originating in transportation systems (e.g., means of transportation and logistics) are among the data which people produce indirectly and unconsciously in physical systems. Until now, such data produced by physical systems was considered, so to speak, as data exhaust [Zikopoulos et al. 2011] of people. Nowadays, however, it is thought that it is possible to recycle such data exhaust and to generate business values out of them.

In the report of the above mentioned researches of IDC, data produced in science, the Internet, and physical systems are collectively called big data.

The features of big data can be summarized as follows:

- The quantity (Volume) of data is extraordinary, as the name denotes.
- The kinds (Variety) of data have expanded into unstructured texts, semi-structured data such as XML, and graphs (i.e., networks).
- As is often the case with Twitter and sensor data streams, the speed (Velocity) at which data are generated is very high.

Therefore, big data are often characterized as V^3 by taking the initial letters of these three terms Volume, Variety, and Velocity. Big data are expected to create not only knowledge in science but also values in various businesses.

By variety, the author of this book means that big data appear in a wide variety of applications. Big data inherently contain “vagueness” such as inconsistency and deficiency. Such vagueness must be resolved in order to obtain quality analysis results. Moreover, a recent survey done in Japan has made it clear that a lot of users have “vague” concerns as to the securities and mechanisms of big data applications. The resolution of such concerns are one of keys to successful diffusion of big data applications. In this sense, V^4 should be used for the characteristics of big data, instead of V^3 .

Social media data are a kind of big data that satisfy these V^4 characteristics as follows: First, sizes of social media are very large, as described in chapter one. Second, tweets consist mainly of texts, Wiki media consist of XML (semi-structured data), and Facebook articles contain photos and movies in addition to texts. Third, the relationships between users of social media, such as Twitter and Facebook, constitute large-scale graphs (networks). Furthermore, the speed of production of tweets is very fast. Social data can also be used in combination with various kinds of big data though they inherently contain contradictions and deficits. As social data include information about individuals, sufficient privacy protection and security management are mandatory.

Techniques and tools used to discover interesting patterns (values) from a large amount of data include data mining such as association rule mining, clustering, and classification. On the other hand, techniques used to mainly predict occurrences of the future, using past data, include data analysis such as multivariate analysis [Kline 2011].

Of course, data mining and data analysis must more and more frequently treat such big data from now on. Therefore, even if data volume increases, data mining algorithms are required to be executable in practical processing time by systems realizing the algorithms. If the processing time of an algorithm increases proportionally as the data volume increases, then the algorithm is said to have linearity with respect to processing time. In other words, linearity means that it is possible that processing time can be maintained within practical limits by some means even if data volume increases. If an algorithm or its implementation can maintain such linearity by certain methods, then the algorithm or implementation is said to have scalability. How to attain scalability is one of the urgent issues for data mining and data analysis.

Approaches to scalability are roughly divided into the following: scale-up and scale-out. The former approach raises the processing capability (i.e., CPU) of the present computers among computing resources. On the