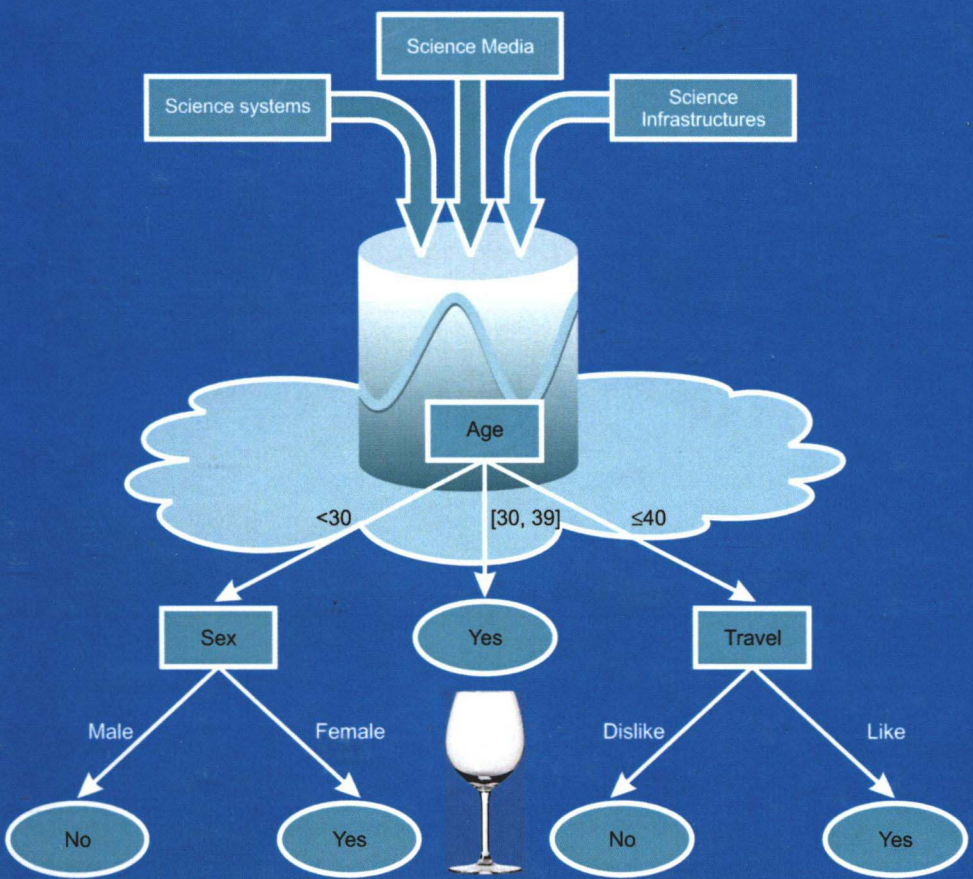
Social Big Data Mining

Hiroshi Ishikawa



CRC Press
Taylor & Francis Group
A SCIENCE PUBLISHERS BOOK

Social Big Data Mining

Hiroshi Ishikawa
Dr. Sci., Prof.
Information and Communication Systems
Faculty of System Design
Tokyo Metropolitan University
Tokyo, Japan

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

Printed on acid-free paper
Version Date: 20150209

International Standard Book Number-13: 978-1-4987-1093-0 (Hardback)

Social Big Data Mining

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.

Hiroshi Ishikawa

July, 2014 Kakio, Dijon and Bayonne

Contents

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).

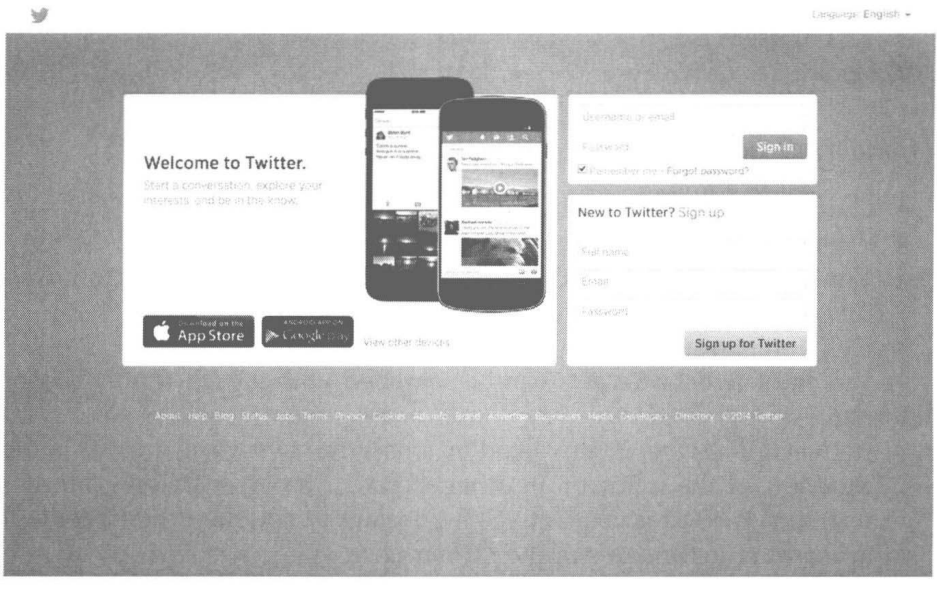


Figure 1.1 Twitter.

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

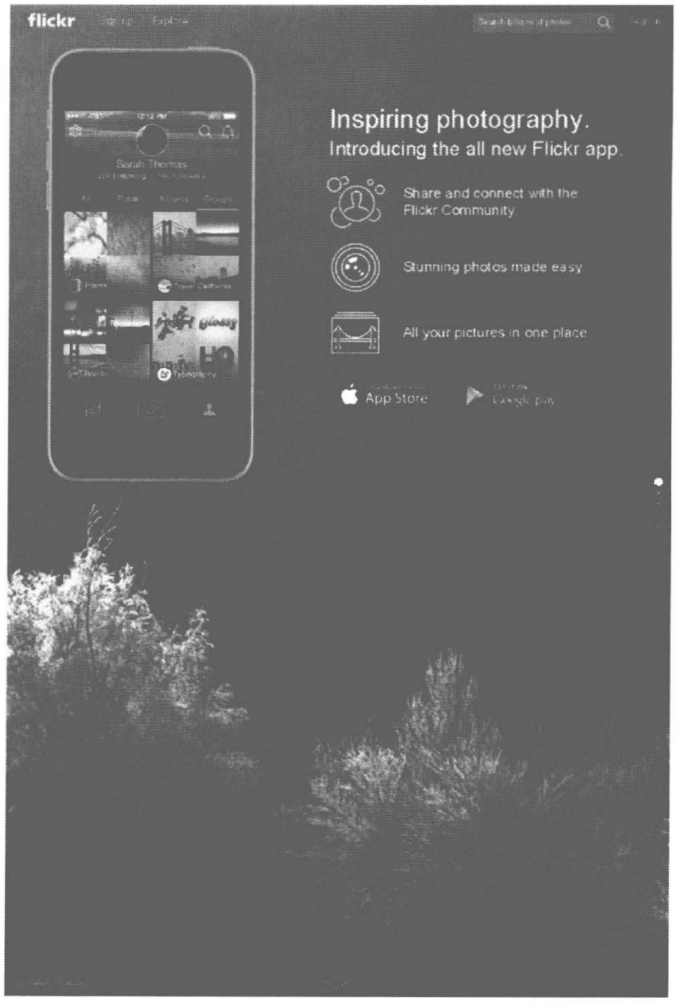


Figure 1.2 Flickr.

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