

ま え が き

現代では、われわれを取り巻くさまざまな分野（科学、インターネット、社会システムなど）で大量のデータが時々刻々と作り出されている。ある調査会社によれば、2011年には人類の生産するデータの総量が、人類が手にする記憶媒体の記憶容量の合計を一桁以上も上回ったという見積りがある。さらに2020年にはデータの生産量は40ゼットバイトになると見積もられている。

対象の観測や分析によって大量データを作り出す科学分野には、宇宙、環境、素粒子物理、生命、医学などが含まれる。今この時点でも電波望遠鏡、人工衛星、粒子加速器、DNAシーケンサ、MRI（核磁気共鳴画像法）装置が、それぞれの分野で絶えず大量のデータを供給し続けている。

科学という専門分野でなくても、一般の人がインターネットを通して意識的に直接作り出すデータには、デジタルライブラリ、およびニュース、Web、Wiki、そしてソーシャルメディアがある。最近になって急速に発展してきたソーシャルメディアの代表としては、ツイッター、フリッカー、フェイスブック、YouTube などがある。また、ウィキペディアもソーシャルメディアの一種ととらえることができる。これらのデータをまとめてソーシャルデータという。

一方で、人間が現実世界における物理的な社会システムの中で無意識のうちに間接的に作り出すデータには、電力（機器）、ガス（機器）、ビル（赤外線センサ、温度、照度、湿度、二酸化炭素などのセンサ、監視カメラ）、輸送（交通機関におけるICカード利用、物流）由来のデータが含まれる。これまでは、社会システムで作り出される、こうしたデータは言わば人間の廃棄データ（data exhaust）と見なされてきた。しかしながら、今日では廃棄データでさえ、リサイクルしてビジネス上の価値をくみ出すことが可能であると考えられている。こうしたデータを実世界データと呼ぶことにする。

すでに言及した先の調査会社によれば、こうしたソーシャルデータや実世界データは、総じてビッグデータと呼ばれている。その特徴は以下のようにまとめられる。

- ビッグデータはその名前のとおり、その量 (Volume) が尋常ではない。
- その種類 (Variety) は、構造データからテキスト、半構造、グラフ（一般にネットワーク）などのように拡大している。
- ツイッター、センサデータストリームなどのように、そのデータが生成される速度 (Velocity) が大きい。

そこでしばしば、ビッグデータの性質は、特徴的なタームの頭文字をとって V^3 (Volume, Variety, Velocity) と言われる。そうしてこれらのビッグデータからは、科学分野の知識に限らず、さまざまなビジネス上の価値が生まれることが期待されている。

ここでは Variety によって、ビッグデータの応用分野が多様であることも併せて意味する。また、ビッグデータは曖昧性 (Vagueness) を多く含み、質のよい結果を得るためにはそうした曖昧性を極力減らす必要がある。さらに利用者の中には、自分のデータを含むビッグデータ応用におけるプライバシーやセキュリティの仕組みに関して漠然とした不安 (Vague concerns) があることも事実である。この不安の払拭が、ビッグデータ応用の普及のカギになると言ってもよい。そこでビッグデータの特徴、4 番目の V (Vagueness) を加えて V^4 と表現してもよいかもしれない。

ビッグデータ時代のデータ分析や応用に携わる者は、しばしばデータサイエンティストと呼ばれ、今や引く手あまたの花形職種の一つとなりつつある。そこでこうしたデータサイエンティストに求められる能力や知識を改めて考えてみる。それには少なくとも以下が含まれる。

- 仮説を立てる
- 仮説を検証する
- ソーシャルマイニング, Web マイニング
- 自然言語処理
- 知識表現

- 可視化
- GIS（地理情報システム）
- 幅広い応用（応用統計学，ビッグデータ応用事例）
- スケーラビリティ
- 倫理と法
- セキュリティ，プライバシー
- コミュニケーション能力（引き出す，理解，説明）

本書ではソーシャルビッグデータに関連する項目に絞って，以下のような基本的な概念や関連する技術を説明する。

- ソーシャルデータとビッグデータ
- 仮説の概念
- 仮説を立てるためのデータマイニング
- 仮説を検証するための多変量解析
- Web マイニング，メディアマイニング
- 自然言語処理
- ビッグデータ応用
- スケーラビリティ

本書では，ソーシャルビッグデータの時代ですますます重要な役割を持つことになった仮説を基本として，ソーシャルビッグデータのための分析技術をとらえるという視点を貫く。そうした意味で本書は類書とは異なり，しっかりとした学術的基盤に立ちつつ，基本から応用まで，ソーシャルビッグデータの全体像を伝えることを目指す。

本書が学生，および若い技術者や研究者をはじめとして，ソーシャルビッグデータに関心のある読者に幅広く利用されることを願う。また本書をまとめるにあたって大変ご協力をいただいた，株式会社コロナ社の方々に深くお礼を申し上げます。

2014 年 9 月

石川 博

目 次

I 編 ソーシャルビッグデータの基本概念と応用

1. ソーシャルメディア各論

1.1	ソーシャルメディアとは何か	1
1.2	代表的なソーシャルメディア	3
1.2.1	ツイッター (Twitter)	3
1.2.2	フリッカー (Flickr)	6
1.2.3	YouTube	8
1.2.4	フェイスブック (Facebook)	10
1.2.5	ウィキペディア (Wikipedia)	11
1.2.6	Web	14
1.2.7	その他のソーシャルメディア	15

2. ビッグデータとソーシャルメディア

2.1	ビッグデータ	17
2.2	実世界とソーシャルメディアのインタラクション	21
2.3	統合的フレームワーク	25
2.4	インタラクションのモデル化と分析	28
2.5	統合分析モデル：概念層	32
2.6	仮説の生成と検証：論理層	37
2.7	もう一つの関心：インタラクションマイニング	40
2.8	分散並列計算基盤：物理層	43

2.8.1 データベースと NoSQL	44
2.8.2 並列分散コンピューティングの仕組み：MapReduce	46
2.8.3 Hadoop	48

3. ビッグデータ時代の仮説

3.1 仮説とは何か	51
3.2 仮説のためのサンプリング	55
3.3 仮説の検証	56
3.4 仮説を立てる	57
3.4.1 帰納的推論	58
3.4.2 演繹的推論	59
3.4.3 もっともらしい推論	60
3.4.4 アブダクション	62
3.4.5 相 関 関 係	63
3.4.6 因 果 関 係	64
3.4.7 類 推	68
3.4.8 推 移 律	69
3.5 仮説の抽象度	70
3.6 仮説再訪	72

4. ソーシャルビッグデータの応用

4.1 分析対象としての Web とソーシャルメディアの相違	74
4.2 構成要素に基づくソーシャルメディア分析の応用	76
4.3 目的から見たソーシャルメディア分析の応用	78
4.4 MiPS モデルによる具体例の記述	81
4.4.1 単 純 な 仮 説	83
4.4.2 より複雑な仮説	88
4.4.3 疑似相関関係を含む仮説	92

4.5 展 望	96
---------------	----

補遺 1. ビッグデータ時代のデータサイエンティストに 求められる能力	97
--	----

Ⅱ 編 ソーシャルビッグデータの分析に用いる技術

5. データマイニングの基本概念

5.1 データマイニングとは	100
5.2 関連技術と問題意識	102
5.3 データマイニングのタスク	106
5.4 基本的なデータ構造	107
5.5 データの質	112
5.6 ま と め	114

6. 相関ルールのマイニング

6.1 応 用	116
6.2 基 本 概 念	118
6.3 相関ルールの種類	120
6.4 アプリオリアルゴリズムの概要	122
6.5 相関ルールの生成	128

7. クラスタリング

7.1 応 用	131
---------------	-----

7.2 データ構造	132
7.3 距離	133
7.4 クラスタリングとアルゴリズム	135
7.5 パーティショニングに基づくクラスタリング	137
7.6 階層的クラスタリング	139
7.7 クラスタリング結果の評価	145

8. 分類

8.1 動機	147
8.2 分類作業	148
8.3 決定木の導出	149
8.4 属性選択のための指標	150
8.5 分類ルールの生成	153
8.6 基本アルゴリズムの拡張	154
8.7 精度の計算	154
8.8 精度の向上	155
8.9 その他のモデル	155

9. 予測

9.1 予測と分類	159
9.2 予測モデル	161
9.2.1 重回帰モデル	161
9.2.2 非線形関数の変換	163
9.2.3 パス解析モデル	163
9.2.4 多重指標モデル	164
9.2.5 因子分析モデル	166
9.2.6 因子の回転	166

9.2.7 共分散構造モデル再訪	168
9.2.8 因子再訪あるいは次元削除	170

10. Web の構造マイニング

10.1 Web マイニング	173
10.2 構造マイニング	174
10.2.1 計 量 書 誌 学	174
10.2.2 引用文献データベースとインパクトファクタ	176
10.2.3 研究者の価値： h -index	179
10.2.4 威 信	181
10.2.5 PageRank	182
10.2.6 HITS	188

11. Web の内容マイニング

11.1 検索エンジン	191
11.1.1 Web ページのクロールリング	193
11.1.2 Web ページのインデクシング	194
11.1.3 Web ページのランキング	198
11.2 情報検索技術	198
11.3 ページの分類	205
11.4 ページのクラスタリング	208
11.5 マイクログログの要約	211

12. Web の履歴マイニングと情報抽出, 深層 Web のマイニング

12.1 Web の履歴マイニング	213
12.1.1 履歴マイニングと推薦	213

12.1.2	アクセスパタンのクラスタリング	217
12.1.3	協調フィルタリングと Web 個別化	218
12.2	情報抽出	220
12.2.1	情報抽出のタスク	220
12.2.2	アプローチ	222
12.3	深層 Web のマイニング	224

13. メディアマイニング

13.1	XML, 木構造, グラフのマイニング	229
13.1.1	XML マイニングの応用	229
13.1.2	XML 構造マイニング	230
13.1.3	XML 内容マイニング	238
13.1.4	より一般的な構造に関するマイニング	240
13.2	マルチメディアマイニング	247
13.3	ストリームマイニング	251
13.3.1	クラスタリング	252
13.3.2	分類	253
13.3.3	頻出アイテムのカウント	254
13.3.4	トレンド分析	255
13.3.5	類似検索	255

14. スケーラビリティと外れ値検出

14.1	相関分析におけるスケーラビリティへの対策	256
14.1.1	メモリを共有しない方式	256
14.1.2	メモリを共有する方式	257
14.2	クラスタリングにおけるスケーラビリティへの対応	258
14.2.1	階層的アプローチに基づくクラスタリング	258
14.2.2	密度の概念に基づくクラスタリング	261

14.2.3	グラフのクラスタリング	265
14.3	分類, およびその他のタスクに対するスケーラビリティ	267
14.4	外れ値検出	269

補遺 2.	媒介中心性: 構造, 内容, 履歴マイニングの あいだの関係についての注意	271
-------	--	-----

引用・参考文献	273
索引	287

I 編 ソーシャルビッグデータの基本概念と応用

1

ソーシャルメディア各論

ソーシャルメディアは、ソーシャルビッグデータの応用にとって不可欠な構成要素である。ここではまずソーシャルメディアとは何かをよりよく理解するために、ソーシャルメディアを複数のカテゴリに分類して、その特徴を説明する。その後で、ソーシャルビッグデータ応用のための分析対象という観点から重要なカテゴリを複数選び、そのカテゴリで代表的なソーシャルメディアを取り上げて、その特徴を説明する。それらの特徴の説明では、関連する統計、ソーシャルメディアの構造や相互作用、類似メディアとの関係に焦点を絞る。

1.1 ソーシャルメディアとは何か

一般にソーシャルメディア（social media）は、プラットフォームとしての Web 上に構築されたシステムとその利用者から構成される。そして利用者は、システムとのあいだで、システムが許した直接的な相互作用を行う。また、基本的に、ある利用者はシステムからだけでなく他の利用者からも識別される。さらに、利用者は複数集まって明示的なコミュニティ、あるいは一般的なネットワークを作る。こうして構築されるネットワークが、いわゆるソーシャルネットワーク（social network）である。利用者は、ソーシャルネットワーク分析の文脈の中では行為者（actor, アクタ）と呼ばれる。利用者はソーシャルメディアシステムと相互作用を行うと同時に、そのネットワークに参加し、あるいはそれを利用しながら、そのメディアが提供するサービスを享受することができる。

より具体的にはソーシャルメディアは、そのサービス内容の違いに基づいて、おおよそ以下のようなカテゴリに分類できる。

- ブログging (blogging)：このカテゴリのサービスでは、利用者は個人的な出来事を含め、あるトピックに関する説明、感想、評価、考え、行動などを日記風におもにテキストで出版する。
- マイクロブログging (micro blogging)：このカテゴリのサービスでは、利用者は一般のブログより短いテキスト（ツイートは最大 140 文字）で、より頻繁にブログを出版する。
- SNS (Social Network Service)：このカテゴリのサービスは、利用者間のソーシャルネットワーク作りそのものを支援する。
- 共有サービス (sharing service)：このカテゴリのサービスは、動画、楽曲、写真、ブックマークなどさまざまなものを利用者間で共有できるようにする。
- ビデオコミュニケーション (video communication)：このカテゴリのサービスでは、利用者間でビデオを利用してチャットや会議を行うことができる。
- ソーシャルサーチ (social search)：このカテゴリのサービスでは、検索結果に対する利用者の好みや意見が、それ以降の検索結果に反映される。あるいは利用者の問い合わせに対して、他の利用者を含めて人々が直接回答してくれる。
- ソーシャルニュース (social news)：このカテゴリのサービスでは、利用者が一次ソースとしてのニュースだけでなく、すでにある好きなニュースを投稿・評価することができる。
- ソーシャルゲーミング (social gaming)：このカテゴリのサービスでは、SNS を基盤として、その利用者間でコミュニケーションを取りながらゲームを行うことができる。
- クラウドソーシング (crowd sourcing)：このカテゴリのサービスでは、利用者が、ある仕事の一部または全部を、他の利用者にアウトソーシ

グすることができる。

- コラボレーション (collaboration)：このカテゴリのサービスは、利用者間の共同作業を支援したり、その共同作業の結果を出版したりする。

1.2 代表的なソーシャルメディア

ここでは、ビッグデータソースの一つとしてソーシャルメディアを分析し、その結果を利用するというソーシャルビッグデータ応用の観点から、現状でのソーシャルメディアのユーザボリュームとインパクトを考慮して、マイクロブローギング、SNS、動画共有、写真共有、コラボレーションの各カテゴリを取り上げる。それらカテゴリから代表的なソーシャルメディアを一つずつ選択し、さらに比較対照として一般 Web を加えて、メディアごとにそのプロフィール（特徴）を説明する。その際には、分析に有効な以下のようなポイントに注目する。

- カテゴリと創設
- 統計
- 構造
- おもな相互作用
- 類似したメディアとの関係
- API (Application Programming Interface)

1.2.1 ツイッター (Twitter)

〔1〕 **カテゴリと創設** ツイッター [Twitter 2014; Twitter-Wikipedia 2014][†] は、2005 年に Jack Dorsey によって創設されたマイクロブローギングのためのプラットフォームサービスである (図 1.1)。それはライブ性が高く、友人間でたがいの状況を把握するのに適するメディアの開発に関するアイデアからスタートした。ツイッターは、2007 年に開催されたサウス・バイ・ウェスト会議 (South by Southwest interactive, SXSWi) で利用者が急増したことがきっかけで注目されるようになった。また、2011 年には日本で宮崎 駿のアニメ『天空の城ラピュタ』がテレビで放映されたときには 1 秒間で 25.088 件のツ

[†] 引用・参考文献の一覧は、巻末にアルファベット順に掲載する。なお、本書に掲載される URL は編集当時のものである。



図 1.1 ツイッター [Twitter 2014]

イート (tweet, ツイッターの記事) が投稿されて話題となった。

〔2〕 統 計

- アクティブな利用者：200M (M：100 万)
- 検索回数 /1 日：1.6 B (B：10 億)
- 記事数 /1 日：400 M

〔3〕 基本的データ構造

(利用者関連)

- アカウント
- プロフィール

(コンテンツ関連)

- ツイート (tweet)

(関係関連)

- Web サイト, ビデオ, 画像へのリンク
- 利用者間のフォロワー-フォロウイー (follower-followee) 関係
- 検索の保存
- 利用者のリスト
- ツイートのブックマーク

〔4〕 おもな相互作用

- アカウントを作成・削除する。
- プロフィールを作成・変更する。
- ツイートを投稿する：フォローしている利用者によって投稿されたツイートがフォロワーのタイムライン上に現れる。
- ツイートを削除する。

- ツイートを検索する：検索語や利用者名でツイートを検索できる。
- リツイート（retweet）する：利用者があるツイートをリツイートすると、そのツイートがフォロワーのタイムラインに現れる。言い換えれば、自分がフォローしている利用者が、あるツイートをリツイートすると、自分のタイムライン上にそのツイートが現れる。
- 返信する：そのツイートを投稿した利用者に返信すると、投稿者と返信者の両方をフォローする人のタイムラインに現れる。
- ダイレクトメッセージを送信する：自分のフォロワーに直接メッセージを送る。
- ツイートに位置情報を付加する。
- ハッシュタグ（hashtag）を付加する：#で始まる文字列で、それを含むツイートを検索できる。ハッシュタグはしばしばあるトピックを示したり、まとまったコミュニティを形成したりする。
- Web ページの URL をツイートに埋め込む。
- ビデオをリンクとして埋め込む。
- 画像をアップロードして共有する。

〔5〕 **類似したメディアとの関係** 一般のプロギングのためのプラットフォーム [WordPress 2014 ; Blogger 2014 など] と同じように、ツイッターの記事はテキストが中心である。ただしツイッターでも、他のメディアへのリンクを含められる。一方でツイッター記事の文字数は、他のメディアに比べて少なく、ツイッター記事が作成される速度も、それらに比べて速いことが特徴である。

なお、WordPress は単にプロギングのプラットフォームとして利用されるだけでなく、LAMP（Linux Apache MySQL PHP）スタック上で比較的容易にシステムの構築ができるので、企業の CMS（Content Management System）としても広く利用されている。

〔6〕 **API** ツイッターは Web サービスの API として、REST（Representational State Transfer, HTTP プロトコルに従って URI とパラメータを指定し

てサービスにアクセスしXML データを得る) とストリーミング (streaming) API を提供している。

1.2.2 フリッカー (Flickr)

〔1〕 カテゴリと創設 フリッカー [Flickr 2014; Flickr-Wikipedia 2014] は, Stewart Butterfield と Caterina Fake によって創設された Ludicorp 社が, 2004 年に立ち上げた画像共有サービスである (図 1.2)。サービス開始当初, 画像共有そのものは多数の利用者間でのチャットの手段の一つという位置付けであった。しかしその後のサービスでは, 逆に画像共有が主目的になり, やがて, もともとの目的であったチャットサービスは問題もあって姿を消した。

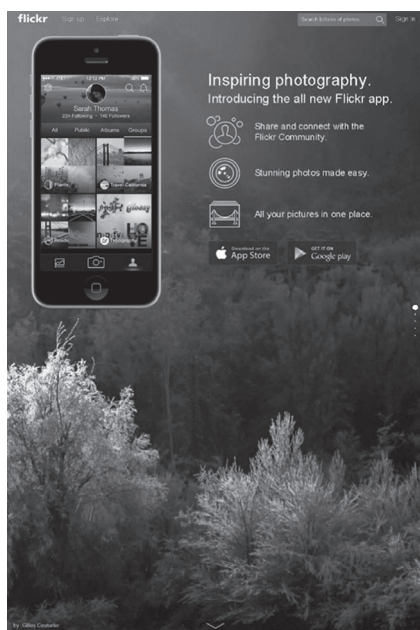


図 1.2 フリッカー
[Flickr 2014]

〔2〕 統計

- 登録利用者: 87M
- 画像数: 6 B

〔3〕 基本的データ構造

索引

【あ】		エントロピー	145	【き】	
アイテム	117	【お】		キー	44
アクセスボタン	217	オーソリティ	188	木構造マイニング	241
アクタ	1	オブジェクト	102	疑似相関	24, 65, 272
圧縮	239	オープンアクセスジャーナル	38, 91	疑似相関関係	92
アブダクション	62	オープンアクセスメディア	38	技術動向調査	79
アプリオリ	122	【か】		キー属性	108
アンカ	196	回帰木	160	帰納的推論	58
アンカテキスト	247	階層的凝集クラスタリング	140	キーバリューストア	44
アンカ伝播	197	階層的分割クラスタリング	144	帰無仮説	56
アンサンブル学習	37, 101	概要抽出	232	逆インデクス	196
【い】		カウント分散	256	客観的観測値	87
威信	181	学習	148	共引用	209
一貫性	47	拡張サーファーマデル	186	教師付き学習	222
因果関係	24, 64, 66	格納スキーマ	236	教師なし学習	223
因子回転	166	可視化	98	協調フィルタリング	220
因子分析	166	仮説	51, 56	共分散構造分析	30, 34
インタラクション	22, 28, 74	仮説検証	31	共分散構造モデル	165, 168
インタラクションマイニング	40, 42	仮説検証的	72	共有サービス	2
インデクシング	192, 194	仮説探索的	72	強連結	184
インデクス	192	仮説立案	29	距離	132
インパクトファクタ	178	カテゴリ	101	【く】	
引用文献データベース	176	カテゴリカル属性	108	空間データ	110
【う】		可用性	47	鎮効果	142
ウィキペディア	11	関心	26, 31, 37, 40	クラウドソーシング	2
ウェーブレット	247	関心度	53	クラス	101
【え】		完全リンク	142	クラスタ	131
影響関係	34	観測変数	34	クラスタ条件	136
影響力	40	関連解析	52	クラスタ分析	101
エクサ	18			クラスタリング	35, 101, 131, 208
演繹的推論	59			グラフ	19, 21, 110, 181
				——の同型	245
				グラフマイニング	244

クローリング 191, 193

【け】

形態素解析 195

計量書誌学 174

系 列 110

結果整合性 47

決定木 149

ゲノム研究 52

検索エンジン 191

検索語 195

検 出 85

検 証 56

検証の因子分析 167

【こ】

構 造 77

構造マイニング 174

構造明確化 239

固有价值問題 183

コラボレーション 3

コンシステントハッシュ法

45

コンテンツ 77

コンテンツ解析 192

コンテンツ獲得 80

コンフィデンス 118

【さ】

最近傍 142

再現率 201

最大固有价值 182

サイバーワールド 42

索引語 195

サーチエンジン 13

サポート 118

サポートベクタマシン

157, 205

三段論法 59

サンプリング 55, 251

【し】

ジオタグ 248

時空間データ 110

シグニチャー 202

時系列 38, 110

次元の呪い 21, 105

実世界データ 21, 42, 94

質的分析 90

重回帰 161

集合知 114, 228

重相関係数 162

従属変数 161

集 約 252

重要度 85

主観的観測値 87

順インデクス 196

順序属性 108

順序変数 86, 135

障害耐性 47

消費行動分析 78

情報抽出 220

情報利得指標 151

人材マネジメント 81

深層 Web 15, 224

信頼度 57, 62

【す】

推移的閉包 210

推移律 69

推 薦 80, 215

数学的帰納法 58

数量化理論 86

スケッチング 251

スケラビリティ

20, 45, 105, 256

スケラブルクラスタリング

91

スケールアウト 20

スケールアップ 20

ストリームマイニング 251

【せ】

正則行列 167

ゼッタ 18

遷移確率 214

線形回帰 161

潜在因子 34

潜在的顧客 41

潜在変数 34, 165

センサデータ 22

戦略マネジメント 81

【そ】

相関関係 63

相関係数 64

相関分析 116

相関ルール 35, 116

属性選択 150

ソーシャルゲーミング 2

ソーシャル広告 10

ソーシャルサーチ 2

ソーシャルデータ 21

ソーシャルニュース 2

ソーシャルネットワーク 1

ソーシャルビッグデータサイ

エンス 27

ソーシャルメディア 1

【た】

大規模可視化 31

ダイナミックな産業構造 96

タイムライン 38

対立仮説 56

多次元尺度構成法 210

多重共線性 164

多重指標 164

タスク 32

多値変数 135

束 127

多変量解析 37

多変量解析ツール 35

単一リンク 142

探索的因子分析 166

単純サーファーマデル 185

【ち】

知の構造化 96

チームワーク 99

直交行列 167

【つ】

ツイッター 3

ツイート 4

強いルール 129

【て】

定性的分析 26

定量的分析 25

適合性フィードバック 201

適合率 201

テキストマイニング 98

データ 98

データ圧縮 239

データウェアハウス 103

データ解析 20

データ行列 109

データ洪水 17

データサイエンティスト 97

データ削減 113

データ収集 30

データ生成 240

データ洗浄 112

データ選択 113

データ統合 113

データベース管理システム

103

データ変換 113

データマイニング

20, 26, 37, 100

【と】

問い合わせ処理 193

統合的フレームワーク 25

統合分析モデル 32

特徴語 199

特徴ベクトル 103

独立変数 161

トピック 85

トランザクション 109

トレンド調査 78

トレンド分析 255

【な・に】

内 積 200

内部的指標 146

ナイーブベイズ 207

ナイーブベイズ法 156

内容検索 247

内容分析 238

ニコニコ 9

【は】

媒 介 66

媒介効果 90

媒介中心性 271

媒介変数 66

廃棄データ 18

パス解析 163

パス係数 164

パースト 38, 212

外れ値検出 269

ボタン 106

ハッシュ木 127

ハッシュタグ 5, 117

ハ ブ 188

パブリックリレーションマ

ネジメント 80

半教師付き学習 223

半構造 21, 102, 229

半構造データ 110

反事実的 67

【ひ】

ピアレビュー 12

被引用回数 178

東日本大震災 39

非周期的 184

非線形関数 163

非線形変数 135

ビッグオブジェクト 29, 32

ビッグデータ 18

ビッグデータサイエンス 21

ビデオコミュニケーション

2

ビュリティ 145

評 価 85

評価極性 39

評価極性分析 90

表 現 98

表層 Web 15

非類似度 132, 141

頻出アイテムセット 119

頻出木 241

頻出グラフ 244

【ふ】

フィードバック分析 80

フェイスブック 10

フォロワー 38, 117

フォロワー 38

フォロワー-フォロワー 4

ブックマーク 15

フリッカー 6

プロギング 2

分 割 131

分割耐性 47

分 類 53, 101, 147, 159

分類器 101

分類タスク 35

【へ】

平均リンク 142

並列 RDBMS 47

べき乗法 182

ベクトル空間モデル

103, 200

偏回帰係数 161

【ほ】

ポスティング 197

ホットスポット 38

【ま】

マイクロプロギング 2

マイクロブログ 211

マクロな仮説 70

マーケットバスケット分析

マーケティング	116	モデル適合度	162	リスク	80
マルチメディアマイニング	79	モニタリング	80	リツイート	5
	247	問題の設定	29	リポジトリ	191
マンハッタン距離	133			流通	79
		【ゆ】		量的予測	90
【み】		有意水準	56	履歴	77
ミクロな仮説	70	ユークリッド距離	133	履歴マイニング	213
ミンコフスキー距離	133			リレーショナルデータベース	
		【よ】		管理システム	45
【め】		要 約	211	リレーション	102
名義属性	108	欲張り法	222	リンク解析	192
名義変数	86	余弦尺度	200	隣接行列	181
メタタグ	196	予 測	159		
メタデータ	111			【る・れ】	
メタ分析モデル	30	【ら】		類似検索	255
		ライフサイクル	52	類似度	133
【も】		ランキング	193, 198	類 推	68
もっともらしい推論	60	【り】		レコード	107
モデル	106	リアルタイムセンサ	92		

【A】		【D】		【H】	
Amazon Mechanical Turk	16	DataGuides	232	h -index	179
AOG	252	DBLP	177	HAC	140
API	3	DBMS	103	Hadoop	48
ArnetMiner	177	DBSCAN	261	HBase	48
		DENCLUE	263	HDC	144
【B】		DHP	257	HDFS	48
Bagging	155			Higgs 粒子	68
Bing	13	【E】		HITS	188
BIRCH	259	ETL	46	Hive	50
		EXIF	39	Hoeffding bound	253
【C】		Exif	248	Holdout 法	154
CAP 定理	47			【I】	
Chameleon	265	【F・G】		IC カード	22
CiNii	178	FQL	11	InnoCentive	16
CMS	5	GIS	98	Instagram	7
CURE	260	Google	13		
		Google Scholar	177		
		GPS	21, 105		

【J・K】

Jaccard 係数	203
k -means 法	137
k -NN	156
KDD	106
k 交差検定法	154
k 最近傍法	206

【L】

LAMP	5
Lance-Williams 係数	143
Lotka の法則	175
LPI	171
LSI	170, 204

【M】

MapReduce	46
MDL	236
MediaWiki	14
MiPS モデル	28, 81

【N】

NoSQL	44
Nupedia	12

N グラム	195
---------	-----

【P】

P 2 P	257
PageRank	104, 182
Pig	50
Pinterest	7
p 値	56

【R・S】

REST	5
Science Citation Index	177
SEM	35
SERP	193
Slashdot	12
SNS	2
SOAP	8
SOM	171
SQL	45
SVM	157

【T】

Tanimoto 係数	203
TFIDF	103, 200

【U・V・W】

USTREAM	9
V^3	19
Ward 法	143, 217
Web	14, 74
Web 個別化	218
Web マイニング	173
Web ロボット	213, 214
Wiki	12
WordPress	5

【X・Y・Z】

XML	19, 229
XML-RPC	7
XML Schema	238
XTRACT	235
YouTube	8
Zipf の法則	175

【数字】

2 乗誤差	137
2 値変数	134
3 層スキーマ	27

—— 著 者 略 歴 ——

1979 年 東京大学理学部情報科学科卒業
1979 年 株式会社富士通研究所勤務
1992 年 博士（理学）（東京大学）
2006 年 静岡大学教授
2013 年 首都大学東京教授
現在に至る

ソーシャル・ビッグデータサイエンス入門

— 基本概念からマイニング技術，応用まで —

Introduction to Social Big Data Science

© Hiroshi Ishikawa 2014

2014 年 11 月 7 日 初版第 1 刷発行



検印省略

著 者	いし かわ ひろし 石 川 博
発 行 者	株式会社 コロナ社
代 表 者	牛 来 真 也
印 刷 所	萩原印刷株式会社

112-0011 東京都文京区千石 4-46-10

発行所 株式会社 コロナ社

CORONA PUBLISHING CO., LTD.

Tokyo Japan

振替 00140-8-14844・電話 (03) 3941-3131 (代)

ホームページ <http://www.coronasha.co.jp>

ISBN 978-4-339-02482-1

(安達)

(製本：愛千製本所)

Printed in Japan



本書のコピー，スキャン，デジタル化等の
無断複製・転載は著作権法上での例外を除
き禁じられております。購入者以外の第三
者による本書の電子データ化及び電子書籍
化は，いかなる場合も認めておりません。

落丁・乱丁本はお取替えいたします