まえがき

ここ数年，データ分析への関心が高まっている．たとえば，ビッグデータという言葉をマスメディアでも見かけるようになった．ビックデータは，その名が示すように大規模なデータのことであるが，データ構造が多種多様であることも特徴の一つである．文章あるいはテキストは，従来の統計的データ分析が対象としにくい構造をもち，そのデータサイズもきわめて大きい．本書でもいくつかの章（論文）でブログやツィッターへの投稿の分析が取り上げられているが，これらは典型的なビックデータであろう．

テキストを対象とした分析技術あるいは分析過程をテキストマイニングと呼ぶ．従来のデータ分析においても，たとえばアンケートにおける自由記述欄のように，文章を収集することは行われてきた．しかし文章を直接あつかおうとすると読解と整理に手間がかかるだけでなく，分析者の主観的判断に左右され，結果を再現することが難しくなる．実際，さまざまな分野で大量のテキスト型データが蓄積されてはいるが，処理手順と客観性の問題から十分に活用されていないのが現状である．

一方，1990 年代後半になると，自然言語処理と呼ばれる分野の技術が一般の PC でも利用可能になり，文章を文字や単語に分解することが容易になった．単語などの単位に分割することで，テキストは分割表や行列の形式にまとめられ，他の数値データと同様の扱いが可能となる．データ処理にはデータマイニングと呼ばれる分野があり，伝統的な統計解析技法に人工知能や機械学習の技法を加え，多様な視点から分析が可能になっている．テキストマイニングは自然言語処理とデータマイニングが融合された新しい分野である．

テキストマイニングが注目を浴びるようになって久しいが，上述のように自然言語処理とデータマイニングの両方に関する知識と技術が必要となるため，実際に導入するにはハードルが高いと感じられているようである．あわせて，現実のデータに対する応用事例も広く知られているとは言いがたい．

本書では，テキストマイニングを実践する研究者らが，それぞれの分野での応用事例をわかりやすく説明している．分野を列挙すれば，金融，医療，言語学，文学，社会調査，政治学，マーケティング，工学，心理学，教育学，宗教学であり，また対象としているデータは，ツィッター，電子メール，自由記述形式のアンケート，文学テキスト，日本語コーパス，新聞記事，議事録，科学論文，学校教材，聖書など多種多様である．それぞれの論文には，研究成果だけではなく，研究のモチベーション，手法の解説，手順の紹介が含まれているので，読者は自身の関心テーマに応用するためのヒントを多く得られるだろう．なお，第 1 章ではテキストマイニングの歴史や理論について総合的に解説がなされている．

残念ながら市販のテキストマイニングソフトは非常に高価である．これもまた，テキストマイニングがデータ分析の選択肢として普及することを妨げている要因と思われる．そこで，無料で利用できるツールと利用方法，さらには応用可能性についての解説を掲載した．無料ではあるが，いずれのツールも有料のソフトに勝るとも劣らない機能が備わっている．本書前半の論文を通じて何か着想を得ることがあれば，ぜひ，これらのツールを活用して分析を実現してほしいと思う．

最後になるが，多忙な中，本書のためにご寄稿くださった執筆者の方々，また企画および編集を通じてさまざまな便宜をはかってくださった共立出版の横田穂波氏にお礼を申し上げたい．

2012 年 10 月 10 日

編　者