巻頭言

世の中にはテキストデータがあふれています．ニュースやブログなど日々膨大な量のテキストが作り出され，スマホやパソコンでアクセスできるようになっています．また，多くの企業では，各種アンケートや顧客対応記録に加え，さまざまな報告書などのテキストデータが作成され，蓄積されています．そういったテキストデータを分析し活用しようという技術がテキストアナリティクスです．その中でも，大量のテキストデータから有益な知見を獲得しようという技術がテキストマイニングです．

データマイニングで分析するデータは，集計しやすい構造になっています．そのため，膨大な量のデータが対象でも，正確な分析結果を期待することができるのがデータマイニングの特長です．しかし最初から集計しやすいように設計された構造化データでは表現しきれない内容がたくさんあります．それを分析対象にできるのが，テキストアナリティクスであり，テキストマイニングです．

例えば，商品レビューで満足度を数値や選択肢で入力する部分のデータは，構造化データであり，集計が容易です．しかし選択肢に含まれない具体的な良さや悪さに関しては，フリーテキストの部分に記述されていても，単純には集計できません．構造化データから一般的な満足度が高い商品や低い商品を特定することはできますが，なぜ満足したか，満足しなかったかを把握するには，フリーコメントのテキスト部分を分析する必要があります．構造化データを対象としたデータマイニングのみでは把握しきれない背景情報や意外な特徴などの分析を実現するのが，テキストアナリティクスやテキストマイニングの特長です．

今や誰でもネット上の膨大なテキストにアクセスできるようになっていますが，このような状況は1990 年ごろから始まったことで，人類の長い歴史の中では，ごく最近のことです．特に，日本語のテキストの電子化と（それをネット上でアクセスできるようにする）オンライン化が本格的に普及したのは 21 世紀に入ってからではないでしょうか．1990 年代には，PC やワープロの普及もあり，テキストデータが量産されるようになってきましたが，基本的にはすべて紙で保管されていました．それが，2000 年問題（西暦の 1999 年を下二桁の 99 だけで表現するコンピュータのシステムが，2000 年になったとたんに誤動作する問題）を避けようと，多くの組織でシステムを新しくした際に，テキストの電子化やオンライン化が加速されたようです．

そうした電子化テキストの普及を背景に，この 20 年程度の間に大きな進化を遂げてきたのがテキストアナリティクスでありテキストマイニングです．この進化のために，日本においてテキストアナリティクスの研究の中心的な役割を担っている学会組織の 1 つが一般社団法人電子情報通信学会の「言語理解とコミュニケーション研究会（通称 NLC 研）」です．NLC 研では，2011 年から毎年「テキストマイニング・シンポジウム」というイベントを主催し，多くの参加者を集めてきました．2017 年 9 月の第 11 回からは「テキストアナリティクス・シンポジウム」と名前を変えて，テキストマイニングやテキストアナリティクスの研究の活性化を図り，最先端の研究を牽引しています．

本書は，この NLC 研の幹事団が執筆しました．監修と第 III 部などの執筆を担当している榊剛史氏

は，2020 年度まで委員長として NLC 研をリードしていました．そして，第 I 部の前半を担当している嶋田和孝氏が 2021 年 10 月時点での委員長です．さらに第 I 部の後半を副委員長の小早川健氏が，第 II 部を副委員長の吉田光男氏と幹事の坂地泰紀氏および幹事の石野亜耶氏が担当しています．私自身，2010 年に NLC 研の委員長に就任し，2011 年の第 1 回テキストマイニング・シンポジウムを企画し，それ以来，NLC 研の専門委員会にずっと参加してきました．こうしたメンバーが中心となって，テキストアナリティクスやテキストマイニングをもっと普及させ，さらに発展させるために，最先端のテキストアナリティクスを活用して成果を出してもらうための本を作ることを目指しました．

集計が容易な数値などの構造化データと異なり，単純に集計できないテキストデータは，構造化データを対象としたデータマイニングの技術だけでは分析できません．テキストデータを分析するためのさまざまな工夫が不可欠です．

例えば，家族構成に関するアンケートを考えてみましょう．家族欄に，父親，母親，配偶者，兄弟といった選択肢があり，父母や配偶者に関しては，いる場合は 1，いない場合は 0 を，兄弟に関しては人数の値を記述するような形式になっていれば，例えば，数十万人や数百万人といった大規模な集団を対象として，各対象者の家族の人数やその平均値，家族構成の分布などを簡単に集計し，分析することができます．これを「あなたの家族に関して自由になるべく網羅的に記述してください．」とフリーテキストで回答してもらうとしましょう．いろいろ書くことができますから，単なる人数だけではなく，何歳離れているか，優しいか，厳しいかなど，多様な情報が記述される可能性があります．その観点から，単なる選択肢のデータよりも情報量が多いと考えられます．しかし，その内容を読み解くのは意外と難しいのです．例えば，「父は亡くなっていません．」と書かれていると，それは，「父親は，すでに亡くなっており，この世にいません．」という意味にも，「父親は，亡くなってはおらず，この世にいます．」という意味にも解釈することができます．人は，基本的に自分の考えに沿った解釈を行う傾向があるため，例えば占い師から「あなたのお父さんは亡くなっていませんね．」と言われると，「なぜわかるのだろう．この人は自分のことを知っている．」と不思議に思えてしまうかもしれません．

日常的に言葉を使っていて気づかないことが多いのですが，言葉から構成されるテキストデータにはさまざまな曖昧性が含まれています．例えば，「長野」という文字列が，「長野さん」という人の名前を示すこともあれば，「長野県」や「長野市」といった地名を示すこともあります．また，人名の場合には，「ながの」と発音したり「ちょうの」と発音したりします．そして，同じ表現で異なる内容を示せるだけでなく，逆に，「長野」を「ながの」「ナガノ」と表現することも含め，同じ内容を異なる表現で示すこともできます．また，「外国人参政権」という文字列を「外国」「人参」「政権」に分けたりしないように注意する必要があります．こういったテキストデータの複雑性がテキストアナリティクスの難しさに繋がります．

テキストアナリティクスで分析したいのはテキストに書かれている内容なのですが，テキストに書かれている内容が確実に正しく解釈できるとは限りません．そもそも，正確な内容が書かれていない可能性もありますし，誤字脱字や文字化けなどで，データが正確に読み取れないこともあります．上記の家族構成に関するアンケートの場合，同居していない祖母や祖父のことは書かない人がいるかもしれませんし，従兄弟姉妹のことまで書く人がいるかもしれません．したがって，構造化され，単純に集計可能なデータの分析とは質が異なることを理解する必要があります．個々のデータの解釈が間

違っている可能性やデータ自体の不正確性など，多様なノイズが含まれていることを前提として分析することが重要です．

そんなノイズだらけのデータが役に立つのかと思われるかもしれませんが，うまく分析すれば確実に役に立つものです．逆に言えば，成果を出すためには，うまく分析する必要があります．ノイズだらけのデータであっても，そこに含まれる内容の分布の偏りや変化を捉えることで有用な気づきを得ることができます．例えば，家族構成に関するアンケートのフリーテキストで，祖母や祖父に関して記述されている割合が高い（すなわち偏りが大きい）地域はどこか，あるいはその割合が大きく変化している地域はどこかといった情報は，データ量が膨大で，その中に含まれているノイズが比較的均一であるなら意味があると考えられます．従来扱えなかったテキストデータを分析対象にすることで，有用な知見が得られる可能性が高まります．

私自身，1990 年代後半からテキストマイニングに取り組み始めて以来，さまざまなテキストデータを分析して，役立つ結果が得られる体験を何度も繰り返してきました．コールセンターの顧客対応記録を分析して，商品の不具合を早期に発見したり，営業成績向上に繋がる知見を獲得したりした際には，ユーザから非常に感謝されましたし，SNS のテキストデータを分析して，発見した飲食店や宿屋で大きな満足感を得たことも数多くあります．そういった嬉しい体験をずっと続けているために，この技術の価値を信じており，知って欲しいという気持ちが強いのです．

前述した通り，ネット上の膨大なテキストに誰でもアクセスできるようになったのは比較的最近のことです．テキストといえば紙に書かれていて人が読むものという考え方がそれまでの長い期間に定着しているため，テキストアナリティクスやテキストマイニングの考え方が理解されにくいと感じることがあります．膨大な量のテキストをすべて読むことはできませんから，それを活用するにはさまざまな工夫が必要です．すべてを理解することは諦め，データから何がわかりそうか，データから何がわかれば役に立つかを考える必要があります．テキストにはさまざまな情報が含まれています．同じ本に関する読書感想文が人によって異なるように，テキストデータから何を読み取るかは人によって異なります．そのため，AI にデータを入れてしまえば何らかの有用な結果が得られるというものではありません．人による工夫の余地が大きいのです．

アイディア次第でさまざまな分析が可能になるのがテキストアナリティクスの面白さです．その反面，多くの場合，簡単に結果が出るものではありません．諦めずに試行錯誤を続けることが重要です．基本的には多様な可能性に思いを巡らせることが有効です．それには経験の蓄積が活きてきますので，やればやるほど成果を出しやすくなります．自分で実際にデータを処理し，試行錯誤をしてみるのがテキストアナリティクスのスキルを向上させる近道です．その考えから，本書では試してみることを重要視しています．

読者の皆さんが，テキストアナリティクスやテキストマイニングを楽しみ，周囲の人に感心され喜んでもらえる成果を出せるようになることを願っています．

2021 年 10 月

那須川 哲哉

日本アイ・ビー・エム株式会社東京基礎研究所主席研究員