第6章

テキストデータにおけるベイジアンネットワーク

本章では、テキストデータにベイジアンネットワークを適用して分析することの可能性と効果について解説します。
テキストデータはあらゆるビジネス領域で活用の場面があり、それを分析することには大きな意義があります。テキストデータを分析する手法には従来テキストマイニングが使われることが一般的ですが、ここにベイジアンネットワークを組み合わせて適用することで分析はより有用なものへと進化します。
本章ではその実現のためのアプローチについて、さまざまな種類のテキストデータにおける適用例や特許文書データを用いた具体的な分析事例を交えながら解説します。

（執筆：株式会社アナリティクスデザインラボ
野守　耕爾）

6.1 はじめに

データをビジネスに活用する取り組みが活発化するなか、文章で記されたテキストデータの活用も進んでいます。定性的なデータであるテキストデータには多くの情報が含まれており、ビジネスの問題解決において、こうしたテキストデータを分析することは、今まで気づかなかったような重要なインサイトの獲得が期待できる取り組みです。

テキストデータを分析する際に用いられるデータマイニング手法であるテキストマイニングという言葉は、1990 年代半ばから世の中に広がったといわれますが [1]、2000 年を過ぎた頃からフリーソフトウェアを含むさまざまなツールが登場し始め、ビジネスの場面でもテキストデータの分析と活用が広く実施されるようになりました。たとえば、株式会社 NTT データ数理システムの Text Mining Studio [2] は、専門知識がなくても GUI の画面上の簡単な操作でテキストマイニングを実行できるツールで、多くの利用実績が生まれています。

テキストマイニングを実行することで、テキストの内容を人間がすべて読み込まなくても、登場する単語の抽出と集計に基づいて全体の記述傾向の現状把握をすることができます。ここにベイジアンネットワークを組み合わせて適用することで、より有用な分析に進化します。特に、テキストデータに潜む要因関係を構造的にモデル化できること、テキストの現状把握だけでなく状況の変化に伴う確率的なシミュレーション（確率推論）ができることという点において、ベイジアンネットワークはその威力を発揮し、ビジネスの問題解決において効果的な施策を検討する強力なツールになりえます。

本章では、テキストデータにベイジアンネットワークを適用して分析することの可能性と効果について、その実現アプローチと適用事例を紹介しながら解説します。6.2 節では一般的に取り組まれているテキストデータの分析と活用について解説し、6.3 節ではそれにベイジアンネットワークを適用することの可能性について解説します。6.4 節ではテキストマイニングとベイジアンネットワークだけでなく、その連携に重要となるトピックモデルも組み合わせて開発した Nomolytics という新しい分析手法を紹介し、6.5 節ではその Nomolytics を適用してテキストデータからモデルを構築し活用する例を、テキストデータの種類ごとに紹介します。6.6 節ではその中でも特許文書データを対象とした実際の分析事例を紹介します。6.7 節では本章の結びとしてテキストデータにベイジアンネットワークを適用することの有用性と、ベイジアンネットワークの使い手が心得ておくべき注意事項をまとめます。

なお、本章で使用する「テキストデータ」とは、「文章情報」と「属性情報」を含むデータとしています。

6.2 テキストデータの分析と活用

本節ではまず、一般的に取り組まれているテキストデータの分析と活用について、特にテキストデータの活用例を挙げながら、テキストデータを分析することの意義とその分析手法となるテキストマイニングの概要について解説します。また、近年の急速な AI の進化によって登場した大規模言語モデルについても紹介し、テキストマイニングとの位置づけを解説します。

6.2.1 テキストデータを分析することの意義

ビジネスにおけるテキストデータの活用例を挙げると、たとえば自由記述回答のあるアンケートのテキストデータを分析すれば、通常の選択式回答からは得られない回答者の潜在的なニーズを把握でき、マーケティングに活かすことができます。口コミと呼ばれる Web 上のレビュー情報のテキストデータを分析すれば、消費者の評価を理解して商品企画に活かすこともできます。また、コールセンターに寄せられた問合せ内容が記録されたテキストデータを分析すれば、顧客の要望や不満を理解して自社の商品やサービスの改善に活かすこともできます。他にも、公開されている特許情報のテキストデータを分析すれば、競合他社の技術動向を把握し、自社の研究開発の計画や M&A などの経営戦略に活かすこともできます。

テキストという定性的なデータは、定量的に収集される他のデータでは得られないような情報を多く含んでいます。たとえば、近年の IoT の技術の進展により、人間の移動や購買などの履歴、生体情報といった人間の行動に関する膨大なセンサデータが収集可能になっていますが、こうしたセンサデータは機械が取得した人間の行動の「結果の情報」であり、その行動に至った「プロセスの情報」は不足しています。一方、人間が自ら文章で記したテキストデータは、人間が自らの行動、あるいは他者の行動をその周辺情報や背景と共に具体的に記録したものもあり、人間の行動を理解する上で有用な情報となることが考えられます。人間の行動発生という結果に対して、なぜそのような行動に至ったのか、そこにどのような要因が関連しているのかといったことが理解できれば、そうした行動を促進あるいは抑制するような施策を講じることもできます。このように、テキストデータを分析して活用することはビジネスの問題解決において有用で、価値の高い取り組みであるといえます。

6.2.2　テキストマイニングという分析手法

テキストデータを分析しようとしたとき、その文章情報は定性的なデータなので、そのままでは他の定量的なデータのように統計解析を実行できません。テキストマイニングはそんな定性的な文章情報のデータを定量的に統計解析可能な形にする自然言語処理技術であり、文章情報に含まれる単語を抽出してその品詞を割り当てる「形態素解析」と、その単語間の文法的な係り受け関係を抽出する「構文解析」を基本技術とする手法です [1]。

たとえば、「函館できれいな夜景を見た」という文章に対して形態素解析を実行すると、「函館（名詞）/ で（助詞）/ きれい（形容動詞）/ な（助動詞）/ 夜景（名詞）/ を（助詞）/ 見（動詞）/ た（助動詞）」となります。なおテキストマイニングのアウトプットでは各単語は原形に変換されて集計されます。さらに構文解析を実行して単語間（正確には文節間）の文法的な係り受け関係を分析すると、「函館⇒見る」「きれい⇒夜景」「夜景⇒見る」という三つの係り受け表現（文法的なつながりのある単語のペア）が抽出されます。

先述の通り、現在では多くのテキストマイニングツールが存在しますが、ほとんどの場合、この形態素解析と構文解析を基本としていることは共通しています。そのため各ツールは、この二つの基本技術をベースにした派生分析機能の種類や結果の可視化の仕方、搭載されている単語の辞書の充実さなどで違いが現れていると筆者は感じます。

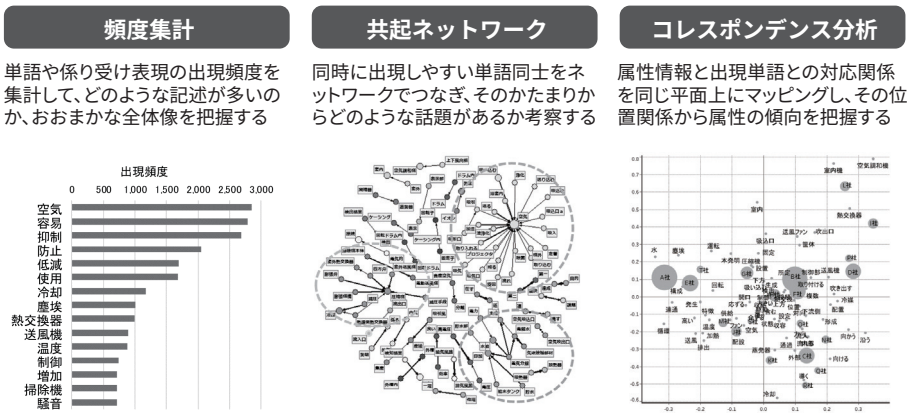


図 6.1　テキストマイニングのアウトプット例

テキストマイニングのツールによくある分析機能のアウトプット例を**図 6.1** に示しました。もっとも単純な分析機能は単語の頻度集計です（図 6.1 左）。形態素解析によって抽出された単語や、構文解析によって抽出された係り受け表現の出現頻度を集計しランキングすることで、

分析対象としているテキストデータの文章情報では、どのような記述が多いのかという全体像を把握します。また、1 件の文章の中で同時に出現しやすい単語どうしをネットワークでつなぐ共起ネットワークという分析機能があります（図 6.1 中央）。単語の頻度集計では、それぞれの単語が独立して集計されるので単語間のつながりがわかりませんが、共起ネットワークではつながりのある単語のかたまりが可視化されます。このネットワーク結果から、このテキストデータの文章ではどのような話題が形成されているのかを考察できます。さらに少し高度な分析機能として、コレスポンデンス分析（あるいは数量化Ⅲ類）というものもあります（図 6.1 右）。これは、テキストデータに紐づく属性情報（たとえばアンケートデータであれば回答者の性別・年代、特許データであれば出願人といった属性情報）と文章情報内の出現単語との対応関係を同じ平面上にマッピングしたものです。そのマップの属性と単語の位置関係からそれぞれの属性の記述傾向を把握できます。

このようにテキストマイニングは、テキストデータの文章情報に含まれる単語や係り受け表現を抽出することを基本とし、その単語や係り受けをベースに集計したり、属性情報とも絡めて統計解析を実行することで、テキストデータ全体の記述傾向を可視化して、現状の特徴を把握できる手法です。

6.2.3　大規模言語モデルの登場

テキストマイニングは従来からある自然言語処理技術を用いたデータマイニング手法ですが、近年では急速な AI 技術の進歩により、大規模言語モデル（Large Language Model, LLM）と呼ばれる革新的な言語モデルが登場しています。これは大量のテキストデータと深層学習を応用したものですが、ここでは代表的な大規模言語モデルについて簡単に紹介します。なお、本章の「テキストデータにおけるベイジアンネットワークの適用」というテーマからは少々離れるので、参考程度の解説として読み飛ばしていただいても問題ありません。

自然言語処理の分野で特にブレイクスルーとなったのは、2017 年に NeurIPS（Neural Information Processing Systems）という AI 分野のトップカンファレンスで Google とトロント大学の研究者によって発表された Transformer という技術です [3]。発表された論文のタイトルが“Attention is All You Need”とあるように、Attention と呼ばれる仕組みだけを用いた深層学習モデルです。

Attention とは、文章の中でどの単語に注目すべきかを判断する仕組みです [4]。たとえば翻訳タスクを考えたとき、翻訳前の入力文章を翻訳後の出力文章に変換する際に、入力のどの単語に注目すべきかを考慮しながら変換し、逆に無関係な情報は無視します。なお、この注目度の重みはニューラルネットワークで学習されます。Attention により、入力と出力の依存関係を捉えることができるため、より精度の高い結果を生成できます。

Transformer は、この Attention だけを用いたニューラルネットワークアーキテクチャであり、従来主流であったリカレントニューラルネットワーク（Recurrent Neural Network, RNN）や畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）を使わないアプローチとして提案されました。また、Transformer は、Self-Attention と呼ばれる自分自身の文章に対して Attention を適用します。もともとの Attention は、入力文章（Source）の単語と出力文章（Target）の単語を対応づけた Source-Target 型の Attention でしたが、Transformer の Self-Attention では、入力文章と出力文章のペアではなく、同一文章の中の各単語が他の単語とどの程度関係しているのかを評価します。これにより単語間の相互作用や依存関係を学習することができます。より具体的な処理としては、各単語の埋め込みベクトルに対して、三つの異なる重み行列を掛けて線形変換を施した、クエリ、キー、バリューという三つのベクトルを作成し、内積を用いた類似度計算を実行することで、他の単語との関連性が考慮された単語ベクトルを獲得する Scaled Dot-Product Attention という計算処理と、その埋め込みベクトルや重み行列などのパラメータを深層学習の誤差逆伝播法によって最適化する学習がされますが、ここでは詳細な理論解説は割愛します。さらに、Transformer では、Multi-Head Attention という、Self-Attention を異なる表現空間（ヘッド）で複数パターン実行する手法により、複数の異なる観点からの類似度を考慮でき、より豊富で柔軟な文脈情報を獲得します。Transformer はもともと機械翻訳タスク向けに開発されたモデルですが、多くの自然言語処理タスクで応用が可能で、どのケースでも非常に高い性能を実現しました。

Transformer は機械翻訳や質問応答など個別のタスクにおいて入力と出力の教師あり学習を行うモデルですが、その後、この Transformer のアーキテクチャをベースにした汎用的な大規模言語モデルが開発されるようになりました。その代表が BERT と GPT です。どちらも大規模なテキストデータセットを用いて、莫大なコストを投じて学習された汎用言語モデルです。従来は個別のタスクごとに学習データを自前で用意し、それを学習する必要がありました。一方、BERT や GPT などは、大規模なデータで言語の汎用的な特徴をあらかじめ学習させておき、その事前学習済みのモデルを再利用することで、後は個別のタスクに応じた比較的小規模なデータを追加学習（ファインチューニング）すれば、高い性能を低コストで実現できるというモデルです。

Transformer は、エンコーダとデコーダの二つのモジュールで構成されていますが、エンコーダでは、入力された文章をベクトル表現に変換する処理を行い、デコーダでは、エンコーダで生成されたベクトル表現から文章を生成する処理を行います。BERT（Bidirectional Encoder Representations from Transformers）は直訳すると「Transformer による双方向のエンコード表現」となりますが、その名の通り Transformer のエンコーダ部分が使われている言語モデルです [5]。一方、GPT（Generative Pre-trained Transformer）は Transformer のデコーダ部分が使われている言語モデルです [6]。BERT の「双方向」という

意味は、Transformer の Self-Attention を実行するときに、各単語は同一文章中の左右にある他のすべての単語との関係性を捉えるということです。ただ、これは通常の Self-Attention の仕組みです。わざわざ「双方向」とつけたのは、先に発表された GPT と対比させた意図があります。GPT は Self-Attention を実行するときに左側にある単語のみが使われ、一方向の関係性を捉えます。ちなみに、GPT は OpenAI によって 2018 年 6 月に発表され、BERT は Google によって 2018 年 10 月にプレプリントという形で発表され、その後 2019 年 6 月に開催された NAACL（North American Chapter of the Association for Computational Linguistics）という自然言語処理分野のトップカンファレンスで発表されました。

BERT では、具体的な処理として、マスクされた言語モデル（Masked Language Model, MLM）という事前学習タスクを実行します。これは、文章中の一部の単語をマスクし、その単語を双方向から予測することにより学習が進みます。つまり大量の穴埋め問題を深層学習で解いているということです。一方、GPT は文章中の単語を左から右に予測していきます。これらの処理は、Attention によって計算された単語どうしの関係性に基づくことで文脈を考慮した推論ができます。こうした処理の特徴により、BERT はより文章全体の文脈理解において優れており、文章分類や感情分析、質問応答（回答抽出）などのタスクに適しているといわれます。一方、GPT は文章生成や文章補完、要約作成、対話生成、言語翻訳などのタスクに適しているといわれます。ただし、これらはその仕組みに基づいた一般的にいわれる得意タスクであり、実際にどのモデルが適しているかは個別のタスクの内容やデータセットに依存します。

ここまで解説した内容、つまり Attention の仕組みだけを用いた Transformer というニューラルネットワークアーキテクチャをベースにした大規模言語モデルの BERT と GPT について、その構成の概要を**図 6.2** にまとめます。

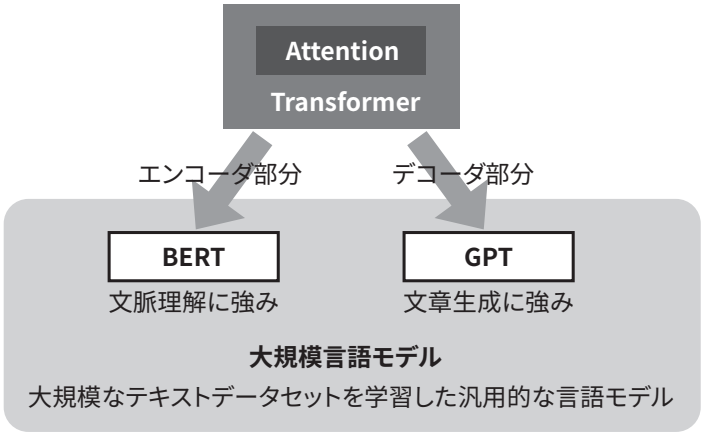


図 6.2　大規模言語モデルの種類と構成の概要

大規模言語モデルをめぐる動向として、BERT は 2019 年 10 月に Google が検索エンジンに採用し、「過去 5 年間で最大の飛躍」と発表され話題となりました。今まで検索キーワードに対して検索結果を返していたところを、文章での検索に対して高い精度の検索結果を返せるようになりました。GPT を開発する OpenAI は 2019 年に GPT-2 を、2020 年に GPT-3 を発表し、2022 年 11 月には GPT-3.5 をベースとした“ChatGPT”という AI チャットサービスをリリースしました。利用者は 2 ヶ月で 1 億人に達し、その性能の高さに世界は騒然としました。2023 年 3 月にはさらに性能が向上した GPT-4 を発表しています。BERT（BERT-Large）ではパラメータ数は 3.4 億個と大規模ですが、GPT は GPT-3 のバージョンでパラメータ数は 1750 億個とさらに超大規模な言語モデルとなっており、その後も超大規模な言語モデルの開発競争が激化しています。

6.2.4　大規模言語モデルとテキストマイニングの位置づけ

こうした大規模言語モデルの登場によって、テキストデータの分析と活用がどのような影響を受けていくのか、ここでは大規模言語モデルと従来のテキストマイニングとの位置づけについて、両者の特徴を対比させながら解説します。

大規模言語モデルとテキストマイニングは、どちらもテキストデータの分析と活用を目的とした手段という意味では共通していますが、それぞれの特徴や用途には大きな違いがあります。大規模言語モデルは大規模なテキストデータセットを事前学習した汎用的な言語モデルであり、その学習された言語の汎用的な特徴に基づいて、文章の文脈を理解したり、新しい文章を生成したりできます。これにより、文章分類や感情分析、文章や要約の作成、対話生成、言語

翻訳などを得意タスクとします。一方、テキストマイニングは手元にある特定のテキストデータの特徴や傾向を理解し、そこからインサイトを得ることを目的とした手法です。特定のテキストデータに出現する単語を抽出し、その単語の頻度情報をベースとした定量的な統計解析を実行することで、図 6.1 で示したような可視化をします。ビジネスの問題解決におけるテキストデータの活用という命題の下では、大規模言語モデルは文章要約や言語翻訳など、そのアウトプットそのものが問題解決において直接的な価値を提供しています。一方、テキストマイニングは、そのアウトプットは問題解決における意思決定に資する価値を提供するものであり、大規模言語モデルと対比させると間接的なアプローチで問題解決に貢献します。

また、それぞれ得意な処理と不得意な処理が存在します。まずは大規模言語モデルが得意で、テキストマイニングが不得意な処理を挙げます。代表的なものとして、大規模言語モデルは文脈を理解することが得意ですが、テキストマイニングはそれが得意ではないということです。たとえば「監督は選手にサインを送った」と「彼は契約書にサインした」という二つの文章で使われる「サイン」は同じ単語ですが、違う文脈で使用されています。テキストマイニングではどちらも同じ「サイン」として認識し、文脈による区別ができませんが、大規模言語モデルではこれらを違う文脈として認識できます。他に、テキストマイニングには感情分析（文章のポジティブ・ネガティブの判定）を行う機能もありますが、これは固定化された語彙リストやルールに基づいて評価するものであり、文脈を捉えた評価になっていません。これに対して大規模言語モデルは文脈を理解した感情分析を実行できます。これらの違いは、テキストマイニングでは特定の単語の出現有無によって文章を捉えますが、大規模言語モデルでは文脈が考慮された単語の特徴ベクトル表現を獲得して文章を捉えるという、処理の違いで現れます。

逆に大規模言語モデルが不得意で、テキストマイニングが得意な処理の例を挙げます。その一つは、大規模言語モデルは大規模なテキストデータセットを事前学習した汎用性のある特徴を捉えたモデルですが、個別性の強い特徴を捉えることは得意ではありません。テキストマイニングでは手元にある特定のテキストデータを処理し、そのデータ特有の個別の特徴や傾向を可視化できます。また、大規模言語モデルのアウトプットは文章生成や質問応答など、その形式は定性的な情報ですが、テキストマイニングのように定量的な統計解析を行うことは得意ではありません。他には、大規模言語モデルは事前学習時にはとてつもない規模のテキストデータを学習しますが、それを再利用するときに入力するテキストデータの規模は限定的で、テキストマイニングのように大量のテキストデータセットを一気に処理することには向いていません。これは大規模言語モデルでは一度に処理できる単語数に制約があるためです。正確には単語ではなく、トークンと呼ばれる単位で処理をしており、大規模言語モデルでは計算コストの爆発抑制のため、このトークンの数に上限を設けています。そのため、大量のテキストデータを一度に処理することを得意としていません。たとえば、BERT の上限トークン数は 512 個、GPT-3 の上限トークン数は 4096 個です。なお、単語とトークンは必ずしも一致せず、トー

クンの数よりも単語の数の方が少なくなることが一般的です。

以上の特徴の違いを観点別にまとめると**表 6.1** のようになります。ただし、この表は大規模言語モデルとテキストマイニングを対比させることで、その違いをわかりやすく理解することを目的に筆者の見解でまとめたものです。タスクの条件や分析データの内容によって一概にはいえないので、あくまでも参考程度に捉えてください。

なお、本章で紹介する分析は従来のテキストマイニングとベイジアンネットワークを組み合わせたアプローチになり、大規模言語モデルは適用していません。これは手元にある大量のテキストデータに潜む特徴や要因関係を把握することを目的とした分析であるためです。

表 6.1　大規模言語モデルとテキストマイニングの特徴の対比

	大規模言語モデル	テキストマイニング
問題解決の方式	**直接的** （問題解決においてタスクそのものが直接的な価値を提供）	**間接的** （問題解決のための意思決定に資する価値を提供）
用途	**文脈の評価や文章の生成**	**テキストデータの特徴可視化**
入力データ規模	**限定的なテキストデータ** （入力単語数に制限あり）	**大量のテキストデータ** （入力単語数に制限なし）
分析の焦点	**文脈** （単語間の相互関係性）	**単語** （特定の単語の出現有無）
結果の形式	**定性的** （テキスト形式による出力）	**定量的** （統計解析による集計と可視化）
活用例・分析例	文章分類、感情分析、文章生成、要約作成、質問応答、対話生成、翻訳	単語頻度集計、単語共起関係把握、話題クラスタリング、属性別傾向把握
適用範囲の性質	**汎用性** （言語の汎用的な特徴を事前学習し多様なタスクに適用可能）	**個別性** （特定のデータセットに存在する個別の特徴や傾向を把握可能）

6.2.5　大規模言語モデルとテキストマイニングの組み合わせ術

こうした違いがある一方で、大規模言語モデルとテキストマイニングを相互補完的に活用できる可能性は大いにあります。たとえば、従来のテキストマイニングの感情分析は精度が高くないことがありましたが、先に大規模言語モデルで文脈に基づいた高い精度の感情分析を施し、各感情別の特徴をテキストマイニングで単語ベースに把握するということが可能です。また、大規模言語モデルでは、特に BERT では文脈に基づいた文章の特徴ベクトルをエンコードできるので、そのベクトル表現に基づいて文章をクラスタリングし、それぞれのクラスターに属する文章群にテキストマイニングを実行すれば、各クラスターの特徴を単語ベースに解釈できます。他にも、多言語のテキストデータをまとめてテキストマイニングしたいときに、大規模

言語モデルで同一言語に翻訳してからテキストマイニングを実行したり、テキストマイニングの辞書作りにおいて、大規模言語モデルを使って類義語の候補を生成するといった活用アイデアも考えられます。

これらの例はすべて大規模言語モデルをテキストマイニングの前処理に活用するものですが、先述の通り、大規模言語モデルは一度に大量のテキストデータを処理することは得意ではありません。対象のテキストデータの量が多いときには、事前に分割してから入力するなどの工夫が必要です。

一方、テキストマイニングを大規模言語モデルの前処理に活用することも有効です。この場合は大量のテキストデータを対象とすることができます。たとえば、まずは従来通りのテキストマイニングの分析を実行することで、特徴を可視化したり文章のクラスタリングをします。この結果から、さらに深掘りして注目すべきデータ対象を絞り込み、その限定された量のテキストデータに対して大規模言語モデルを適用し、感情分析や要約生成などを行えば、その対象の文章理解をより深めることができます。

このように、大規模言語モデルとテキストマイニングを組み合わせることでビジネスの問題解決においてより有用なアプローチを形成できることが考えられ、今後こうした取り組みも盛んに検討されることが予想されます。

6.3 テキストデータの分析に適用するベイジアンネットワークの可能性

本節では、テキストデータの分析にベイジアンネットワークを適用する考え方について解説します。

6.3.1 テキストマイニングにベイジアンネットワークを適用するメリット

テキストマイニングはテキストデータの定性的な文章情報を定量的に統計解析可能な形にする手法であると先ほど説明をしました。統計解析可能な形というのは、テキストマイニングによって図 6.3 に示すようなデータにテキストデータが変換されるということです。テキストマイニングではテキストデータの文章情報に出現する単語を抽出するので、各文章においてその単語の出現頻度をカウントしたデータ（あるいは単語の出現有無を 1 と 0 で示したフラグデータ）ができあがります。こうしたデータ形式を一般的に Bag-of-Words とよびます。テキストデータがテキストマイニングによって統計解析可能になるのは、こうしたデータ形式に変換されるためであり、図 6.1 に示した頻度集計や共起ネットワーク、コレスポンデンス分析といっ

テキストデータ（例：ホテルの口コミ）					文章情報で出現する単語																
ID	性別	年代	評点	コメント	部屋	風呂	広い	快適	朝食	美味しい	綺麗	清掃	スタッフ	対応	丁寧	挨拶	利便性	コンビニ	近い	…	良い
1	男性	40代	5点	部屋が広くて快適に過ごせましたし、朝食がとても美味しかったです。	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0		0
2	女性	20代	4点	部屋の中は綺麗に清掃されていて、お風呂も広くて快適でした。	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0		0
3	女性	40代	4点	帰り際に清掃のスタッフの方が笑顔で挨拶してくれて気持ち良かったです。	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0		1
4	男性	30代	3点	駅やコンビニなどが近くにあって利便性が良く使いやすかったです。	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1		1
5	女性	50代	5点	スタッフの対応が良く、近くのコンビニの場所も丁寧に教えてくれました。	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1		1
…																					

図 6.3　テキストマイニングにより変換されたテキストデータの出現単語データ

たテキストマイニングの各種分析機能は、すべてこのデータ形式をベースにして成り立っています。

このようにデータが統計解析可能な形式になれば、ベイジアンネットワークも適用できます。ベイジアンネットワークは、複数の変数の間にある確率統計的な関係性をネットワーク構造でモデル化し、ある変数の状態を条件として与えたときの他の変数の起こり得る確率を推論できるモデリング手法です。モデルとは、データに潜む特徴や傾向を抽象化したものですが、特にベイジアンネットワークは回帰分析や決定木分析、ニューラルネットワークなどの他のモデリング手法と異なり、目的変数と説明変数の区別がないことが特徴的です。目的変数と説明変数が区別されている通常のモデリング手法では、一つの目的変数ごとにモデルを構築する必要があり、構築されたモデルを使った推論の実行では、説明変数群から一つの目的変数を推論するという一方向の推論に限定されます。一方、ベイジアンネットワークは、目的変数と説明変数の区別がないことによって、変数どうしのお互いの関連性を把握でき、また一つのモデルで複数の推論対象を指定でき、その推論条件の変数も区別なく自由に設定できるので、さまざまな方向から確率推論のシミュレーションを実行できる手法になっています。

テキストマイニングにベイジアンネットワークを適用することの主なメリットを二つ挙げると、①テキストデータに潜む要因関係を構造的にモデル化できること、②現状把握だけでなく状況の変化に伴うシミュレーションができること、です。以下にそれぞれのメリットについて解説します。

①テキストデータに潜む要因関係を構造的にモデル化できる

テキストマイニングでは、文章情報でどのような単語が多く使われる傾向があるのかがわかります。さらにベイジアンネットワークを適用することで、そうした文章情報の傾向がどのような要因によって影響を受けているのか、あるいは他の事象に影響を与えているのかという関係性を構造的なモデルとして可視化できます。たとえば、図 6.3 はホテルの口コミというテキストデータを例に取り上げていますが、口コミには性別や年代、評点といった属性情報があるため、こうした属性情報と口コミのコメントの間に存在する統計的な関係性を分析できます。

②現状把握だけでなく状況の変化に伴うシミュレーションができる

テキストマイニングは基本的に文章情報の記述傾向の現状把握をする手法なので、その現状から状況が変化したときに、それに伴って結果がどう変化し、影響を受けるのかということはシミュレーションしません。これにベイジアンネットワークを適用することで、その確率的なシミュレーションが可能になります。①で述べたように、ベイジアンネットワークを適用することで、テキストデータの文章情報を中心とした要因関係を構造的にモデル化できるようになります。このモデルを利用することで、たとえば性別・年代が変わると口コミのコメントの内

容がどの程度変化するのか、コメントの内容が変わると口コミの評点がどの程度変化するのかというように、与えた条件に対する結果の挙動を確率的にシミュレーションできます。こうした分析によって、たとえばターゲット顧客層の関心に応じたプロモーションを設計したり、顧客満足度を高めるためのサービス改善などを検討できます。つまり、ベイジアンネットワークで構築されたモデルを用いることで、実施する施策を条件として与えたときの効果を評価したり、その効果を最大化させる条件を探索でき、効果的なビジネスアクションを検討できます。これはビジネスの問題解決においてとても強力なツールになります。

6.3.2 テキストマイニング×ベイジアンネットワークのモデル化の方法と限界

ここからはテキストマイニングにベイジアンネットワークを適用して構造的なモデルを構築する具体的な方法を紹介していきます。テキストマイニングで得られた図 6.3 のようなデータにベイジアンネットワークを適用してモデルを構築することは、抽出された各単語を、その出現有無を示す 1 と 0 の状態をもつ確率変数として扱えば可能です。テキストデータの各属性情報も確率変数として扱えば、単語と属性情報の関係をモデル化できます。たとえば「快適」という単語が現れると「評点」が 5 点となる確率が何%となるかといった関係を示すモデルや、「30 代」だと「利便性」という単語を含むコメントをする確率が何%となるのかといった関係を示すモデルを構築できるということです。こうした方法によってテキストデータの文章情報に関連した要因関係を構造的にモデル化できます。

しかし、この方法は、テキストマイニングによって大量の単語が抽出されるほどモデルが複雑になっていきます。たとえば**図 6.4** は病院で収集された 4238 件の子どもの傷害事故の診療記録データ（転落や火傷、誤飲といった事故の情報と受傷した子どもの情報が記録されたデータで、傷害の原因となった製品や子どもの行動といった傷害発生の詳しい状況などは文章で記されたデータ）を対象に、ベイジアンネットワークでモデルを構築した事例です [7]。

この事例では、まず傷害発生の詳しい状況が記された文章情報にテキストマイニングを実行し、その文章に含まれる「製品」に関する単語と、子どもの「行動」に関する単語を抽出します。それらを 1 と 0 の状態をもつ確率変数とし、また属性情報として記録されている子どもの「年齢」「性別」「事故の種類」もそれぞれ確率変数として、それらの関係をベイジアンネットワークで構造的にモデル化しています。これによって各年齢の子どもはどのような製品でどのような行動をとる傾向があり、それによってどのような事故の傷害に至る傾向があるのかという関係性をモデルで示すことができ、さらにそのモデルを用いることで、そうした行動や事故の発生確率をシミュレーションできます。

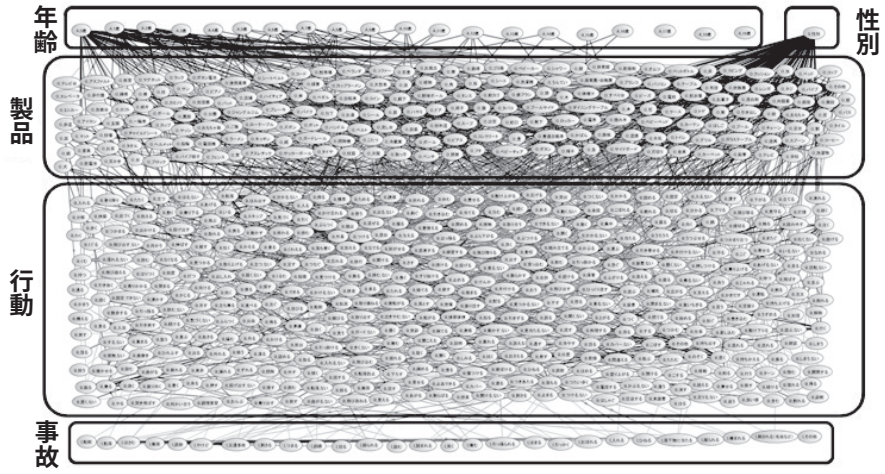


図 6.4　テキストマイニング×ベイジアンネットワークのモデルの例（参考文献 [7] より転載）

たとえば、「コイン」という製品では「飲む」という行動が 57.3%で、「誤飲」という事故が 32.6%と確率が高いことがシミュレーションできます。また「ミニカー」という製品では「押す」という行動が 50.0%、「乗る」という行動が 23.8%で、「転倒」という事故が 31.4%、「転落」という事故が 27.1%と確率が高く、さらにこれらは 1 歳児で確率がより高くなることがシミュレーションでわかります。このモデルは子どもの情報や製品の情報から子どもの起こり得る行動と事故を予測して、子どもの傷害事故の予防を支援するツールとして筆者が開発したもので、2010 年の人工知能学会論文誌に掲載された事例です。

このようにテキストマイニングによって抽出された各単語を確率変数とし、他の属性情報との関係をベイジアンネットワークによって構造的にモデル化すれば、確かにテキストデータに潜む要因関係を示すモデルを得られ、そのモデルを使った確率的なシミュレーションを実行できます。しかし、図 6.4 を見てわかる通り、抽出された単語一つひとつを確率変数とした構造モデルは大変複雑になり、解釈可能なモデルとは言い難いものです。図 6.4 のモデル構築に用いたデータは 4238 件でした。この取り組みを実施していた 2009 年ごろはまだビッグデータという言葉は使われていなかったものの、当時では 4000 件程度でもとても大規模なデータであったため、論文のタイトルを「大規模傷害テキストデータに基づいた～」としているくらいです。「ビッグデータ」という言葉は、2012 年 3 月に当時のアメリカのオバマ大統領が“Big Data Research and Development Initiative”を発表してから一般的に使われ始めたといえますが、今では 4000 件程度でビッグデータという言葉を使う人はいないでしょう。数万件、数十万件、数百万件のテキストデータが解析の対象となっている時代です。単語一つひとつを確率変数としてベイジアンネットワークのモデルを構築する場合、4000 件程度のテキストデー

タですら図 6.4 のような複雑なモデルとなるため、それを大きく超える規模のテキストデータを対象とする場合は、このアプローチには限界があるということは自明の理です。

先述の通り、モデルとはデータに潜む特徴や傾向を抽象化したものであるので、情報がより集約されて全体としてシンプルに表現されていることが望ましいものです。ビッグデータとなるテキストデータを対象とする場合は、テキストマイニングによって得られた図 6.3 のようなデータをそのままベイジアンネットワークのインプットにするのではなく、その前に情報を集約する一工夫が求められるという発想が浮かびますが、その方法を次節で紹介します。

6.3.3　トピックモデルのクラスタリングを応用したモデル化

単語一つひとつを確率変数にしてベイジアンネットワークのモデルを構築するとモデルが大変複雑になるのであれば、同様の意味をもつ単語群をグルーピングして、その単語グループを確率変数とすれば、この問題は解決できるという発想は自然かと思います。人間が目視で確認してグルーピングをしてもよいですが、人間の作業では主観的になり、担当者によってもルールが変わることもあり、また大量の単語を分類する作業の負荷はあまりにも大きいです。そこで機械的なグルーピングを実現するために、データを分類するクラスタリング手法を適用します。

データクラスタリングの手法では、階層クラスター分析の Ward 法や非階層の k-means 法などが有名です。これらの手法は、基本的に要素間の距離に基づいてクラスターを形成しますが、要素の数がとても多い場合には、要素間の距離が自然と離れて妥当な結果が得られにくい次元の呪いという問題が起きてしまいます。したがって、要素数の多いデータ、つまり列数の多い高次元のデータのクラスタリングには適していません。テキストデータの分析では図 6.3 のように、基本的に単語一つひとつを列に取った超高次元のデータとなるので、これらの手法はテキストデータのクラスタリングに向いていないということがわかります。

テキストデータを対象としたクラスタリングには、自然言語処理や人工知能学の分野で開発されたトピックモデルと呼ばれる手法があります。トピックモデルは共起行列と呼ばれる行列データをインプットとし、行の要素 x と列の要素 y の背後にある共通する特徴となる潜在クラス z を抽出する手法です。テキストデータのクラスタリングに用いる場合、この潜在クラスをトピックと呼びます。たとえば図 6.3 のようなデータ形式は、文章 x（行）×単語 y（列）という形式の共起行列となり、各文章情報に対する単語の共起頻度がデータ化されています。これにトピックモデルを適用することで、各単語の出現情報を学習し、使われ方の似ている単語群で構成される潜在的なトピックを抽出し、そのトピックを軸に行要素の文章情報を分類できます。

図 6.3 のテキストデータはトピックモデルを適用することで**図 6.5** のように変換されます。

トピックは各文章の所属確率（図 6.5 上）と各単語の所属確率（図 6.5 下）が計算される形で抽出されます。また、Ward 法や k-means 法などの従来のクラスター分析では、列の情報に基づいて行の情報をクラスタリングするか、あるいは行の情報に基づいて列の情報をクラスタリングするというように、クラスタリングの対象は行か列どちらか一方になりますが、トピックモデルでは行と列を同時にクラスタリングできます。つまり抽出された同じトピックに対して、行要素の文章と列要素の単語が同時に所属し、それぞれに対して合計 100%となるように所属確率が割り振られることになります。またトピックモデルは、大量の単語を列とする超高次元データに対して、いくつかのトピックを列とする低次元データに変換できる手法なので、次元圧縮法とも呼ばれます。

テキストデータ（例：ホテルの口コミ）					トピック				
ID	性別	年代	評点	コメント	部屋の広さ	朝食	清潔さ	スタッフ対応	利便性
1	男性	40代	5点	部屋が広くて快適に過ごせましたし、朝食がとても美味しかったです。	0.09%	0.08%	0.04%	0.01%	0.02%
2	女性	20代	4点	部屋の中は綺麗に清掃されていて、お風呂も広くて快適でした。	0.08%	0.01%	0.09%	0.05%	0.01%
3	女性	40代	4点	帰り際に清掃のスタッフの方が笑顔で挨拶してくれて気持ち良かったです。	0.02%	0.02%	0.07%	0.08%	0.02%
4	男性	30代	3点	駅やコンビニなどが近くにあって利便性が良く使いやすかったです。	0.01%	0.03%	0.01%	0.02%	0.09%
5	女性	50代	5点	スタッフの対応が良く、近くのコンビニの場所も丁寧に教えてくれました。	0.01%	0.02%	0.01%	0.09%	0.06%
…									

トピックの構成

部屋の広さ	
単語	確率
広い	21%
部屋	16%
快適	12%
風呂	7%
…	

朝食	
単語	確率
朝食	38%
美味しい	16%
豊富	8%
デザート	4%
…	

清潔さ	
単語	確率
部屋	17%
綺麗	15%
清潔	11%
快適	9%
…	

スタッフ対応	
単語	確率
スタッフ	30%
対応	18%
丁寧	12%
挨拶	6%
…	

利便性	
単語	確率
アクセス	24%
近い	18%
コンビニ	13%
利便性	5%
…	

図 6.5　トピックモデルを適用したテキストデータと抽出したトピック

代表的なトピックモデルの手法としては、潜在意味解析（Latent Semantic Analysis, LSA）[8]、確率的潜在意味解析（Probabilistic Latent Semantic Analysis, PLSA）[9]、潜在的ディリクレ配分法（Latent Dirichlet Allocation, LDA）[10] などがあります。LSA は特異値分解によってトピックを抽出し、LSA を確率的に処理し発展させたものが PLSA です。LSA における特異値分解の行列表記を PLSA では確率モデル（aspect モデル）で表記しま

す。LSA は入力する行列の成分をそのまま使用すると、潜在クラスが大きな値をとりやすいベクトルに引っ張られて抽出される傾向があるため、TF-IDF などで重み付けされた行列を用いられることが多いですが、PLSA はそうした重み付けの事前処理をすることなく潜在クラスを抽出できます。一方、LDA は PLSA をさらに拡張させた手法で、個々の文章情報における各トピックの現れやすさを表す確率が、PLSA ではあくまでも学習させた観測データのみから定義されますが、LDA ではディリクレ分布という確率分布を事前分布に仮定して生成させます。PLSA では観測データに過剰に適合して新規のデータの適合度が下がってしまうオーバーフィッティングが生じやすく、新しいテキストデータにおけるトピックの生成確率は定義されませんが、LDA ではこれを推定できます。こうした特徴の違いもあって、トピックモデルを適用するときは PLSA よりも汎化性能の高い LDA を用いることの方が主流になっているようです。

こうしたトピックモデルによって得られた図 6.5 のようなデータを用いて、その抽出されたトピックを確率変数として扱い、ベイジアンネットワークで構造的にモデル化することで、単語一つひとつを確率変数としてモデル化する場合と比較し、モデルがとてもシンプルになり、全体的な要因関係を解釈しやすくなります。次節では、テキストマイニング×トピックモデル×ベイジアンネットワークの組み合わせでテキストデータに潜む要因関係をモデル化する新しい分析手法について紹介します。

6.4 テキストマイニング×トピックモデル×ベイジアンネットワークによるテキストデータの分析手法：Nomolytics

本節では、テキストマイニング、トピックモデル、ベイジアンネットワークを連携して適用し、テキストデータからそこに潜む要因関係の特徴や傾向を構造的にモデル化する手法として、筆者が開発したNomolytics（Narrative Orchestration Modeling Analytics）[11]を紹介します。

なお本手法は筆者が有限責任監査法人トーマツに所属していたときに特許として出願し登録されたものであり（特許第6085888号）[12]、令和5年5月25日現在で有限責任監査法人トーマツと株式会社アナリティクスデザインラボが権利を保有しているものです。

6.4.1 Nomolytics の分析手法の概要

Nomolytics の分析手法について**図 6.6** に概要を示しながら説明します。本手法では、まずテキストマイニングによりテキストデータの文章情報から単語を抽出し、各単語の共起頻度をデータ化した共起行列を作成します。次に、その共起行列をインプットにトピックモデルのPLSA を適用し、使われ方の似ている単語をトピックにまとめ上げ、全テキストデータに対する各トピックの該当度を計算します。最後に、そのトピックを確率変数として扱い、ベイジアンネットワークによってトピック間あるいは他の属性情報との間の確率的な要因関係を構造的にモデル化します。

本手法ではトピックモデルとして PLSA を採用しています。その理由は、本手法は分析対象としているテキストデータの特徴を率直に理解するための分析だからです。先述の通り、

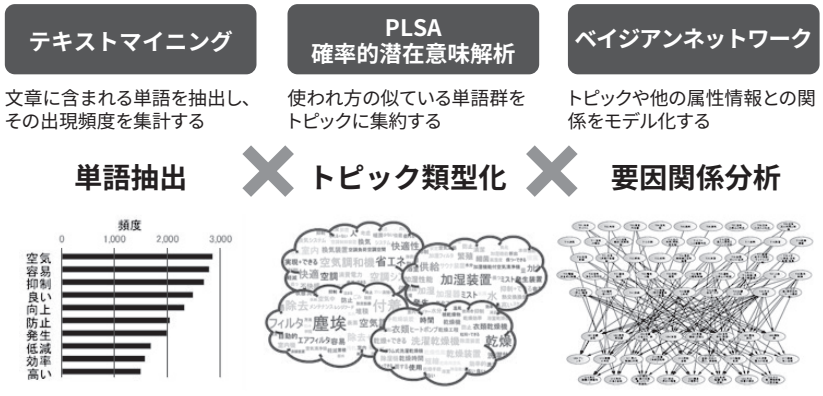
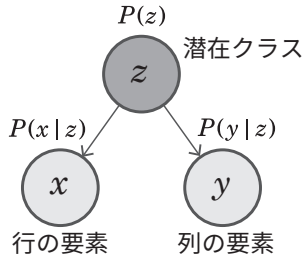


図 6.6　Nomolytics の分析手法

PLSA は観測データにオーバーフィットし、新しいデータの対応が難しいですが、逆にいえば観測データのみからそのデータが真に示す潜在クラスを抽出できるということです。LDA ではディリクレ分布を事前分布に仮定していることで、オーバーフィッティングを回避した汎化性能の高いトピックモデルですが、その分、一般的で抽象度が高い結果となりやすく、観測情報の特徴を純粋に表現したものとはいえなくなります。テキストデータを利用してビジネスの問題解決のための有用なアクションを講じるには、観測データのありのままの特徴を人間が理解することが重要であると考え、Nomolytics では PLSA をあえて採用しています。PLSA の理論の概要については次節で解説します。

6.4.2　PLSA の理論の概要

PLSA の概念図を**図 6.7** に示します。先述の通り、PLSA は共起行列データをインプットとし、行の要素 x と列の要素 y の背後にある共通する特徴となる潜在クラス z を抽出する手法です。これを文章分類で用いる場合は、図 6.3 に示したデータのように、各文章の出現単語を記録した文章 x（行）×単語 y（列）という共起行列を学習し、文章 x とそこに出現する単語 y の間に共通するトピック z を抽出する、ということになります。



x と y の共起確率を潜在クラス z を使って表現する

$$P(x,y)=\sum_{z}P(x\mid z)P(y\mid z)P(z)$$

図 6.7　PLSA のグラフィカルモデル

以下に、トピック抽出までの計算過程を説明します。PLSA では文章 x と単語 y の共起確率 $P(x,y)$ を潜在クラス z を用いて式 (6,1) のように分解して考えます。ここで、文章 x における単語 y の出現回数を $N(x,y)$ とすると、式 (6.2) の対数尤度を最大にする $P(x\mid z)$、$P(y\mid z)$、$P(z)$ を、EM アルゴリズムを用いて計算します。つまり式 (6.3) の E ステップと式 (6.4)〜(6.6) の M ステップを計算することで最尤推定します。以上から PLSA の実行によって得られるアウトプットは 3 種類の確率変数 $P(x\mid z)$、$P(y\mid z)$、$P(z)$ の値になります。図 6.5 の例でみると、図 6.5

上の確率は $P(x|z)$ が該当し、トピック z に対する各文章 x の所属確率を示しています。図 6.5 下の確率は $P(y|z)$ が該当し、トピック z に対する各単語 y の所属確率を示しています。なお PLSA の実行には、たとえば、株式会社 NTT データ数理システムのデータマイニングツール Alkano [13] に搭載されている「二項ソフトクラスタリング」という分析手法を利用できます。厳密には PLSA とは異なる部分がありますが、PLSA と同様に x と y の共起確率 $P(x,y)$ を式 (6.1) のように展開して潜在クラス z を抽出する手法であり、PLSA を代替する類似手法として実行できます。

$$P(x,y)=\sum_{z}P(x|z)P(y|z)P(z) \tag{6.1}$$

$$L=\sum_{x}\sum_{y}N(x,y)\log P(x,y) \tag{6.2}$$

$$P(z|x,y)=\frac{P(x|z)P(y|z)P(z)}{\sum_{z}P(x|z)P(y|z)P(z)} \tag{6.3}$$

$$P(x|z)=\frac{\sum_{y}N(x,y)P(z|x,y)}{\sum_{x}\sum_{y}N(x,y)P(z|x,y)} \tag{6.4}$$

$$P(y|z)=\frac{\sum_{x}N(x,y)P(z|x,y)}{\sum_{x}\sum_{y}N(x,y)P(z|x,y)} \tag{6.5}$$

$$P(z)=\frac{\sum_{x}\sum_{y}N(x,y)P(z|x,y)}{\sum_{x}\sum_{y}\sum_{z}N(x,y)P(z|x,y)} \tag{6.6}$$

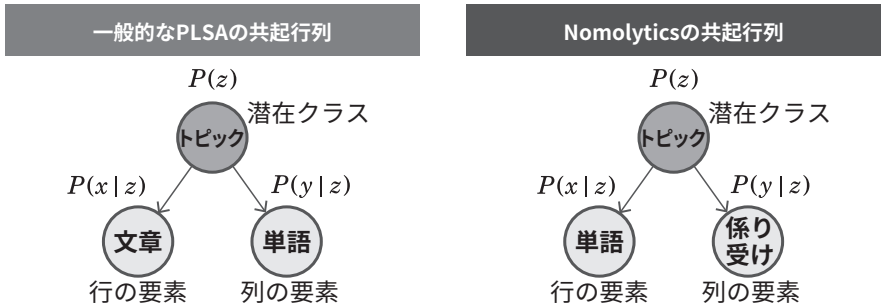
6.4.3 Nomolytics の各手法の連携における工夫

Nomolytics では、より意味性の強いトピックを抽出するため、PLSA のインプットとする共起行列のデータは、一般的なトピックモデルで用いられるものとは異なる行列構成の工夫を施します。ここではその工夫の概要を紹介します。

先述の通り、トピックモデルは行の要素 x と列の要素 y の背後にある、共通する特徴となる潜在クラス z（トピック）を抽出する手法ですが、潜在クラス z には行の要素と列の要素の両方が所属する結果となります。このことから、先ほどは行と列を同時にクラスタリングできる手法という表現をしました。この特性から、共起行列の行と列は双方が十分意味をもつ情報で構成すれば、抽出された潜在クラスの意味を二つの軸から解釈することができ、結果の解釈性が高まりますし、二つの情報軸が集約された意味性の強い結果を得ることも期待できます。

図 6.8 左に示すように、一般的な PLSA の適用では、「文章」×「単語」という構成の共起行列をインプットにしますが、行に設定された「文章」は文章 ID であり、それ自体に意味はもちません。そのため、抽出されたトピックの意味を解釈するときには基本的に「単語」の軸

のみで解釈することになります。また、「文章」×「単語」の共起行列は各単語の出現頻度をカウントしたデータで、1 と 0 で構成されることが多く、そのほとんどは 0 となるスパース（疎）なデータとなります。これでは文章間・単語間で差が出にくく、特徴がクリアなトピックが得られにくいという傾向もあります。



	単語1	単語2	単語3	…
文章ID:1	1	0	1	
文章ID:2	0	0	0	
文章ID:3	0	0	0	
文章ID:4	1	1	1	
…				

	係り受けa	係り受けb	係り受けc	…
単語1	325	264	11	
単語2	241	201	6	
単語3	28	41	288	
単語4	9	15	4	
…				

図 6.8　Nomolytics における PLSA の共起行列構成の工夫

そこで Nomolytics では、共起行列の構成によって解釈のしやすいトピックを抽出する工夫を施します。たとえば、「名詞の単語」×「動詞の単語」というように、行と列をそれぞれ異なる品詞の単語で構成する方法や、図 6.8 右のように「単語」×「係り受け」という構成で、軸の一方を係り受け表現（文法的なつながりのある単語のペア）とする方法をとります。こうした「単語」×「係り受け」の共起行列に PLSA を適用することで、単語という話題の観点となる軸に基づき、その観点の具体的な内容となる係り受け表現をグルーピングでき、より文脈上近しい言葉・表現でまとめられた解釈のしやすいトピックを抽出できることが期待できます。また、こうした共起行列は文章に対する単語の出現頻度をカウントした 0 ばかりのデータではなく、各単語と各係り受けが同時に出現する共起頻度が入るクロス集計型の行列です。こうした行列形式では頻度の値が入りやすくなり、スパース性の問題の影響を受けにくく、頻度の大小の差も生まれやすいため、よりまとまりのあるクリアなトピックを抽出できることが期待できます。さらにその共起行列のサイズは一般的な PLSA で用いる共起行列に比べて（特に行数において）とても小さいものになり、計算時間も大幅に削減できる効果もあります。

また Nomolytics では抽出されたトピックを確率変数として扱い、ベイジアンネットワークを適用することでトピック周辺に存在する要因関係を構造的にモデル化しますが、トピック

を確率変数として扱うための変換処理にも工夫があります。処理の詳細は 6.6 節で解説していますが、ここでは考え方を説明します。「文章」×「単語」という共起行列をインプットとする一般的な PLSA で抽出されるトピックは、所属確率という値によって文章情報にトピックが紐づいた形で結果が得られるので（すなわち図 6.5 のようなデータがアウトプットされるので）、その所属確率という連続値をカテゴリ化処理すれば、トピックは元のテキストデータにおける確率変数としてそのまま扱うことができます。一方、Nomolytics の「単語」×「係り受け」の共起行列で抽出されたトピックは、各単語と各係り受けには紐づいていますが（それぞれトピックに対する所属確率が計算されますが）、元のテキストデータには紐づけがされていない結果となるので、この紐づけを計算する処理が必要になります。そこで、テキストデータの文章情報に含まれる単語・係り受けと、その単語・係り受けが各トピックに対してもつ所属確率から、そのテキストデータに対する各トピックの該当度を示すスコアを確率的に計算します。最終的にはそのスコアに閾値を設け、各トピックに該当するか否かを 1 と 0 のフラグに変換したトピックの確率変数を作成します。

以上のように Nomolytics では、テキストマイニング、トピックモデル（PLSA）、ベイジアンネットワークという三つの技術を連携して適用し、テキストデータに潜む要因関係の特徴や傾向を構造的にモデル化して理解する新しい分析手法です。特にその三つの技術を連携するステップ（共起行列の構成とトピックの確率変数化）で独自の工夫を施していることも手法の特徴となっています。こうした三つの技術を組み合わせることで、膨大なテキストデータの文章情報をいくつかのトピックという人間が理解しやすい形に整理でき、さらにベイジアンネットワークによって、そのトピック周辺に潜む要因関係を構造的にモデル化できます。そしてそのベイジアンネットワークのモデルを用いることで、ある変数の条件を変化させたときに、それに伴って他の変数がどのように変化するのかという確率的なシミュレーションを実行できます。

6.5 テキストデータにベイジアンネットワークを適用したモデル構築例

上記で紹介した Nomolytics のように、テキストマイニング、トピックモデル、ベイジアンネットワークを連携して適用することで、テキストデータに潜む要因関係の特徴や傾向を構造的にモデル化できますが、本節では、テキストデータの種類ごとにそのモデル化のアプローチとその活用例をいくつか紹介します。なお、本節の内容は筆者のこれまでの研究活動や企業のコンサルティング業務を通して適用した事例に基づいています。

6.5.1 自由記述付アンケートデータへの適用

アンケートデータの統計解析はよくある試みですが、選択式の設問と自由記述式の設問が両方用意されたアンケートの場合、従来は選択式の回答データのみ統計解析を実行し、自由記述式の回答データはそれだけ独立して目視で確認するか、テキストマイニングでその記述概要を把握するという、選択式回答と自由記述式回答を別々に分析する傾向にあります。その理由は、選択式の回答は構造化されたデータであるのに対し、自由記述式の回答は構造化されていない定性データであり、そのままでは同時にまとめて統計解析を実行できないためです。アンケートにおいて、選択式の設問はアンケート設計者の仮説に基づいて選択の回答を導くものであり、その回答は受動的といえますが、自由記述式の回答は能動的な回答であり、回答者の生の声の情報であるため、とても貴重な情報です。統計解析のしづらさから、その貴重な回答者の生の声が分析対象から除外されたり、そのデータだけ分離して他の選択式回答の結果との関係が分析されないということは大変もったいない話です。

そこで、自由記述式の回答内容にテキストマイニングとトピックモデルを適用して、その内容をトピックに変換し、アンケートの各回答データに対するトピックの該当度を計算します。そのようにすると、各トピックも他の選択式回答と同様に構造化されたデータとして扱うことができ、選択式と自由記述式の回答を一緒に分析できます。そのトピックの該当度を確率変数として扱い、ベイジアンネットワークを適用すれば、アンケートの各選択式回答の結果と自由記述式回答のトピックの関係構造をモデル化できます。

たとえば、顧客満足度に関するアンケートデータを対象に上記のような分析を実行すれば、**図 6.9** に示すような顧客満足に関わる各設問の関係構造を、自由記述の回答トピックとともにモデル化できます。こうしたモデルを用いることで、顧客は自社の商品に対してどのような期待や価値を感じ、それが満足度やロイヤリティにどのように影響を与えているのかということ

を、自由記述のトピックの要因も絡めて把握でき、顧客理解をより深めることができます。またベイジアンネットワークの確率推論の機能を実行すれば、高い顧客満足やロイヤリティの確率を高めるにはどのような価値を提供することが効果的なのかという推論もできます。なお、図 6.9 のモデル図は各設問タイプを一つのノードにまとめて簡略的に図示したものですが、実際のアンケートデータのモデル化では、設問ごとにノードを設定、あるいは複数の設問を因子分析した設問因子ごとにノードを設定してモデルを構築することになります。

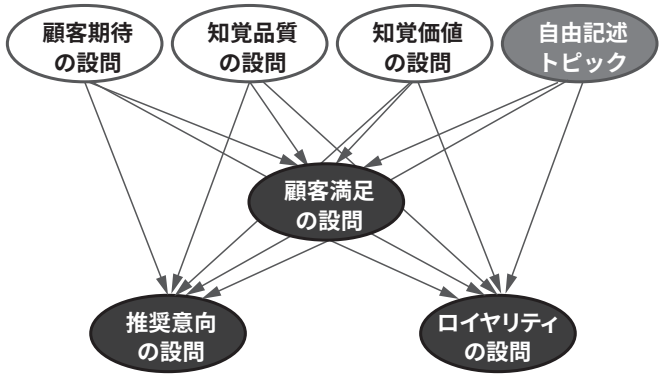


図 6.9　自由記述アンケートデータを用いた顧客満足の構造モデル

また、**図 6.10** に示すように、自由記述の回答から得られたトピックに対する各設問との関係構造をモデル化することもできます。こうしたモデルを用いることで、たとえば回答者の属性（性別や年代、職業など）と、評価対象の商品の属性によって、自由記述の内容がどのように変化するのかということを把握できます。そしてベイジアンネットワークの確率推論を実行すれば、まだ市場に投入していない企画中の新しい商品であっても、その商品の属性情報とターゲット顧客の属性情報から、どのようなコメントが寄せられるのかということを事前に評価できます。こうした構造のモデルを構築できるのは、ベイジアンネットワークが目的変数と説明変数の区別をしないモデリング手法であるためで、複数の目的変数（ここでいう自由記述の各トピックなど）に対する要因の関係構造をモデル化できます。

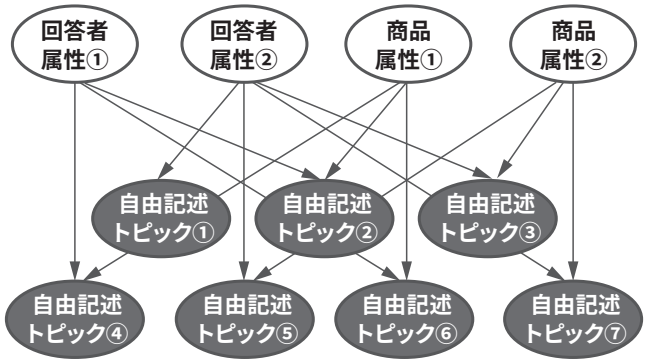


図 6.10　自由記述アンケートデータを用いた自由記述内容の構造モデル

6.5.2　ユーザレビューデータへの適用

ユーザレビューは、消費者によって Web 上に投稿された商品やサービスの評価に関する感想文形式のコメント情報であり、口コミとも呼ばれます。今では宿泊施設やレストラン、観光地、家電製品、化粧品など、インターネット予約や EC サイトの普及により、さまざまなジャンルの投稿情報が大量に蓄積され、Web 上で閲覧できます。

レビューデータはコメントの情報と評点の情報がメインであり、性別や年代といった投稿者の属性情報が取得できるものもありますが、取得できないものもあります。アンケートデータと比較するとレビューデータは取得できる属性情報が限られており、またコメントの内容も丁寧に書かれているものもあれば雑なものや酷い言葉が使われているもの、正しい文法で書かれていないものがあるなど、データの質の面ではレビューデータはアンケートデータよりも劣るかもしれません。しかし、データの量の面ではとても多くのデータを得ることができ、特に不特定多数の消費者の生の声をこれだけ得ることはなかなかできません。また、自社の商品だけでなく、他社の商品に対する評価についてもデータを得ることができるため、自社に限らない業界全体における消費者の評価の傾向を把握できます。企業でもこうしたレビューデータを収集し、自社商品の改善やマーケティング、競合他社分析などに活用する試みがされています。なお、レビューデータは Web 上で誰でも閲覧できますが、そのデータの取得と利用に関してはルールが設定されていることがあるため、利用を検討する際には必ずその Web サイトの利用規約を確認するようにしてください。

レビューデータを分析する場合、集計しやすい評点を集計したり、レビューのコメントを目視で 1 件 1 件確認したり、あるいはテキストマイニングでコメントの全体像を単語ベースに可視化して把握するといったことは従来から実施されています。そこにトピックモデルとベイジアンネットワークを適用することで、コメントから抽出されたトピックとレビューの評点や

投稿者の属性情報などの関係構造をモデル化できます。

たとえば、**図 6.11** に示すように、レビュー評点に対する投稿者の性別や年代、商品属性、レビュートピックの関係をモデル化できます。こうしたモデルを用いることで、消費者の属性や商品の属性、さらにはコメントの内容から評点に影響を与える要因を把握できます。また、ベイジアンネットワークの確率推論によって、各要因の条件から評点の確率分布を推論したり、評点を高めるにはどの要因をどのような状態にすればよいのかという推論をすることもできます。つまり、どのような属性の人がどのような属性の商品にどのようなコメントをすると評価にどの程度影響するのかということを定量的に把握できます。これによって満足度を向上させるためには属性別にどのような商品・サービスが提供されることが望ましいのか、といったことを検討するマーケティング戦略などへの活用が考えられます。

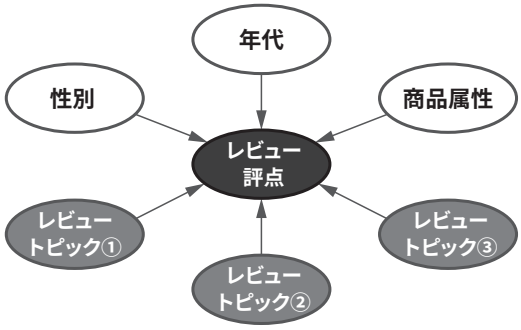


図 6.11　ユーザレビューデータを用いた評点の構造モデル

また、先述のアンケートデータのモデル化と同様に、**図 6.12** に示すような、レビューのトピックに対する各属性との関係構造をモデル化できます。こうしたモデルを用いることで、たとえば消費者の属性と商品の属性によって、レビューのコメント内容がどのように変化するのかということを把握できます。そしてベイジアンネットワークの確率推論によって、商品の属性情報とターゲット顧客の属性情報を条件として入力すれば、どのようなコメントを発する傾向にあるのか、すなわち商品のどのようなことに関心が強いのかということを把握できます。これによって、それぞれの属性条件の関心対象に応じた商品企画をするといった活用が考えられます。

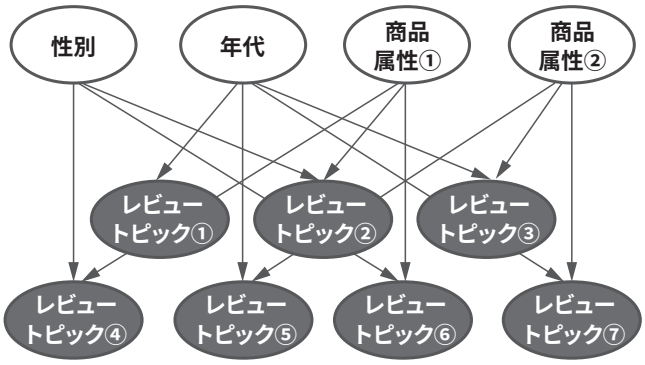


図 6.12　ユーザレビューデータを用いたレビュー内容の構造モデル

実際に、宿泊施設のレビューデータと観光地のレビューデータを用いて上記のようなモデル化を検討した事例を紹介します。宿泊施設のレビューデータを用いたモデル化では、京都府にある宿泊施設のレビューデータ 11535 件を用いて、各項目評点（風呂の評価、部屋の評価、総合評価など）に対するレビュートピックとの関係を、**図 6.13** に示すようにモデル化しました [14]。このモデルを用いたベイジアンネットワークの確率推論結果から、たとえば総合評点が満点となる確率を向上させる要因の一つとして、スタッフの対応に関するトピックが影響していました。そのトピックの内容にはスタッフの対応が丁寧で親切であるということの他に、笑顔が素敵であること、挨拶が気持ちよいこと、嫌な顔をしないことなどのコメントが反映されており、スタッフの表情や挨拶も宿泊サービスにおいて顧客満足に影響するということが示唆されました。

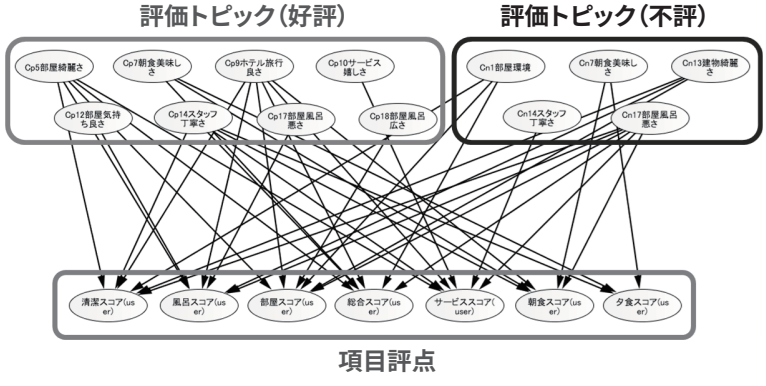


図 6.13　宿泊施設のレビューデータを用いた項目評点に対するトピックの関係モデル（参考文献 [14] より転載）

また、観光地のレビューデータを用いたモデル化では、国内の温泉地におけるレビューデータ 12564 件を用いて、評点（特に 5 点満点中 4.5 点以上という高い評点を付けるか否か）に対するレビュートピックと観光客属性（性別・年代・同行者）との関係を**図 6.14** に示すようにモデル化しました [15]。このモデルを用いたベイジアンネットワークの確率推論結果から、たとえば男性はカップル・夫婦で行く温泉旅行で高評点の確率が上がり、特に温泉で温まることが話題にされるとより有効であることがわかりました。一方で、女性はカップル・夫婦とは行かない（友達や家族と行く）温泉旅行で高評点の確率が上がり、特に砂湯で楽しむことが話題にされるとより有効であるという結果が得られました。つまり、属性の条件によって温泉旅行の満足度を高める要因が変化するということが示されました。

こうした結果を活用することで、温泉という同じ観光資源であっても、ターゲット属性によって異なる魅力を、その属性が接触しやすいメディアで紹介するなど、効果的なプロモーション戦略を検討できます。図 6.14 のモデルの結果を活用すれば、たとえば男性が手に取るような雑誌では、恋人との温泉旅行を提案し、特に寒いなか露天風呂で温まるということを写真を通じて PR します。一方で女性が手に取るような雑誌では、女友達との温泉旅行を提案し、特に砂湯を楽しむ姿の写真を通じて PR するといった検討ができます。

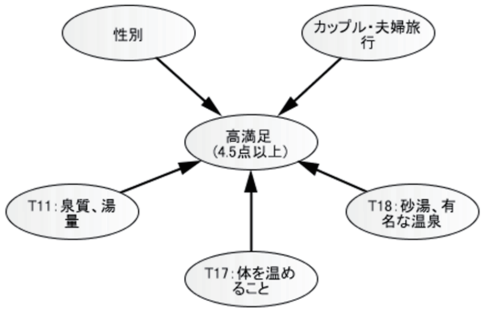


図 6.14　観光地のレビューデータを用いた評点とトピックの関係モデル
（参考文献 [15] より転載）

6.5.3　コールセンターの問合せ履歴データへの適用

企業のコールセンター窓口に連絡のあった問合せ情報は、オペレータがその内容を文章で記録していることが多く、最近では音声から直接テキストに変換されるシステムを導入している企業もあります。問合せデータの記録対象は企業によってさまざまですが、その問合せが企業のどの商品・サービスに対するものなのかに紐づいて記録されていることも多く、取得可能な

場合は問合せをした顧客の属性情報も紐づけられています。また、問合せ内容によってエスカレーション（オペレータがその場での対応が困難な問合せに対して専任者に引き継ぎをすること）の発生有無が記録されていたり、問合せが苦情の場合はその不満度がオペレータの主観で点数付けがされることもあります。他に、契約商品の解約の要求や会員の退会の要求といった、問合せに関する顧客の要求行動の情報が付与されることもあります。

企業のコールセンター部門では、顧客の問合せに対応することがメインの業務であり、その問合せ内容と対応の履歴をデータとして記録するのは、あくまでも顧客対応のためであり、そのデータを分析し、他の業務に活用することを主目的にしたものではありません。しかし、ここで記録・蓄積されているデータは、顧客の生の声による潜在的なニーズが詰まった貴重なデータです。こうした顧客の問合せデータを分析して他の事業部門が活用することは、より顧客目線でビジネスを展開する上でとても価値があるといえます。実際にコールセンターの問合せデータの分析は、企業がテキストマイニングを実施する際によくあるテーマですが、やはりテキストマイニングのツールを使って、従来の単語ベースの可視化に留まるケースが多いといえます。しかし、コールセンターの問合せデータは、企業の規模によっては年間数十万件、数百万件というデータが蓄積されており、企業のテキストマイニングのテーマの中ではトップクラスのビッグデータです。

単語ベースの可視化をして全体像を把握する従来のテキストマイニングでは、その対象がビッグデータだとマイニングで抽出される単語も膨大になるため、結果はとても複雑化し解釈困難なアウトプットになってしまいます。こうした大規模なテキストデータの分析には特にトピックモデルの適用が有効で、大量に抽出される単語をいくつかのトピックに変換することで、従来の単語ベースではないトピックベースで特徴の全体像をシンプルにわかりやすく解釈できます。さらにベイジアンネットワークを適用すれば、顧客の問合せ内容に関連した要因関係を構造的にモデル化でき、顧客ニーズのより深い洞察を定量的に得ることができます。

ここではコールセンターの問合せデータにトピックモデルとベイジアンネットワークを適用してモデルを構築するアプローチの例を紹介しますが、基本的には先述したアンケートデータのモデル化やユーザレビューデータのモデル化と同様です。たとえば、各問合せデータに対してエスカレーションの発生や不満度といった問合せの状態に関する情報があったり、契約の解約要求や会員の退会要求といった顧客の要求行動に関する情報があれば、その結果に対して顧客属性や商品属性、問合せ内容のトピックといった要因の関係構造をベイジアンネットワークでモデル化することで、**図 6.15** に示すようなモデルを構築できます。こうしたモデルを用いることで、どのような属性の顧客がどのような属性の商品にどのような問合せをすると解約につながるのか、あるいはエスカレーションのような対応困難となるのか、あるいは不満度が高まるのかといった、各要因との関係性を把握できます。そしてベイジアンネットワークの確率推論を実行することで、こうした要因から解約確率を推論したり、その解約確率を減少させる

ためにどのような内容の問合せを解消すべきなのかということも推論できるため、顧客の離反防止策の検討などへの活用が考えられます。

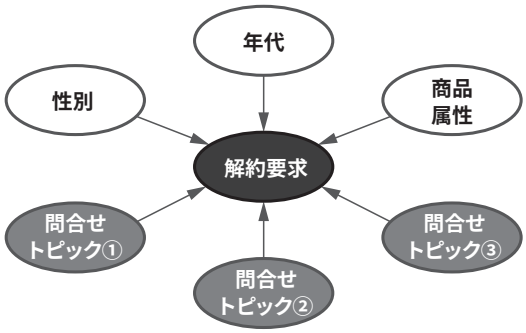


図 6.15　コールセンターの問合せデータを用いた顧客の要求行動の構造モデル

また、先ほどのアンケートデータやユーザレビューデータと同様に、**図 6.16** に示すような問合せのトピックに対する顧客属性や商品属性の関係構造をモデル化することもできます。こうしたモデルを用いることで、どのような属性の顧客はどのような商品でどのような問合せをする傾向にあるのか、その確率を推論できます。この結果を用いることで既存商品の改善に活用することもできますし、新規商品を市場に投入する際には、その商品の属性からどのような問合せが発生する可能性が高いのか事前に予測することもできるので、問合せ対応の事前準備などにも活用することが考えられます。

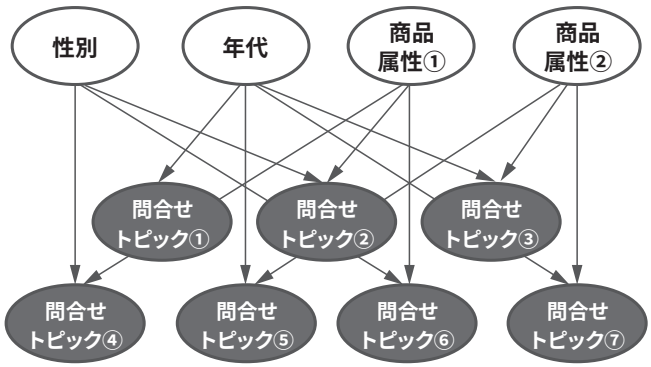


図 6.16　コールセンターの問合せデータを用いた問合せ内容の構造モデル

6.5.4　特許文書データへの適用

テキストデータにベイジアンネットワークを適用してモデルを構築する最後の例として、特許文書データへの適用を紹介します。

企業の技術戦略を検討するうえで、その技術領域の動向を把握するために特許情報はよく分析の対象とされます。特に、通常は他社の技術開発動向は機密性が高いため外部から確認することは難しいですが、特許情報はそれを探ることのできる貴重な公開情報であり、それを分析する価値が高いことは明らかです。特許情報分析というと従来パテントマップ [16] と呼ばれるものが代表的であり、これは主に特許の出願人や出願年、特許分類（IPC、FI、F タームなど）を軸にして特許件数を集計し、出願動向を可視化する分析です。

近年では特許の要約文や請求項、明細書などのテキストデータを分析対象とし、テキストマイニングによって、人間ではなかなか読み切れない膨大な特許文書の内容の全体像を把握するアプローチも採用されています [17]。そのアプローチの一つとして、用途と技術の関係を把握する分析があります。これは、国内で出願された特許では、その要約文には「課題」と「解決手段」という二つの項目が分かれて記述される傾向があり（ルールではなくあくまでも暗黙の記述形式ですが）、その記述形式を利用して分析します。「課題」と「解決手段」のそれぞれの項目の文章をテキストマイニングして単語を抽出し、それを分析者がカテゴリに分類し、**図 6.17** に示すような課題カテゴリと解決手段カテゴリに該当する特許件数をクロス集計します。これによって用途と技術の対応関係を把握し、たとえば、ある用途を実現するための解決技術の候補を探ったり、自社技術の新しい用途展開を探索するといった活用がされます。

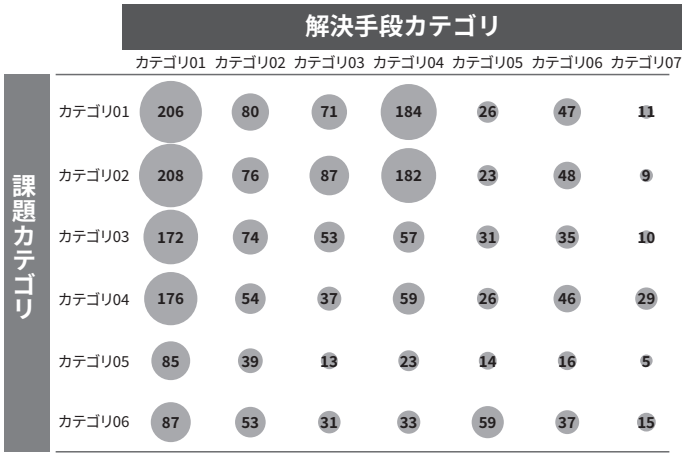


図 6.17　特許の課題文章と解決手段文章をカテゴリ化してクロス集計した分析例

しかし、こうした従来の分析アプローチは、分析対象とする特許データがビッグデータ化するとなかなか難しくなります。先述の通りビッグデータに対するテキストマイニングでは、マイニングによって抽出される単語が膨大になるため、それを分析者が目視でカテゴリ化するのは大変困難ですし、カテゴリ化のルールも属人的になってしまいます。また従来のアプローチでは、単純なクロス集計によって用途と技術の関係性を可視化していますが、統計的な関係の分析はされていません。クロス集計の頻度の大きさだけでは一見関係がありそうな用途と技術でも、それが統計的に意味のある関係であるとは限りません。たとえばその用途や技術に該当する特許のもともとの件数が多ければ当然クロス集計の頻度も大きくなるため、単純なクロス集計では関係性の考察を誤る可能性があります。

こうした従来の分析の問題に対処するアプローチとして、トピックモデルとベイジアンネットワークを適用することは有効に働きます。まず、「課題」と「解決手段」のそれぞれの項目の文章にテキストマイニングとトピックモデルを適用することで、用途に関するトピックと技術に関するトピックを機械的に抽出します。そして、そのトピックを確率変数として扱い、ベイジアンネットワークの適用によって、**図 6.18** に示すような用途トピックと技術トピックの関係構造をモデル化することで、用途と技術の確率統計的な関係性を把握できます。次節では、このアプローチについて具体的な分析事例を紹介します。

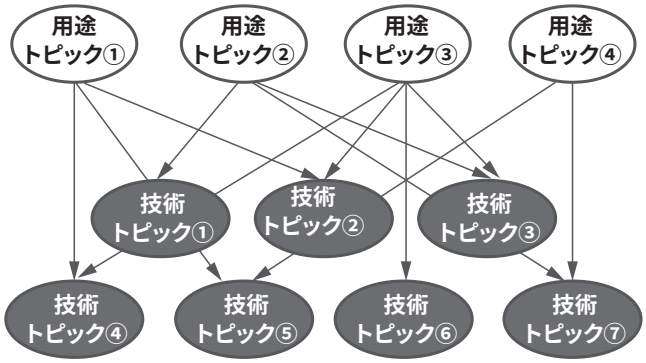


図 6.18　特許文書データを用いた用途と技術の関係モデル

6.6 特許文書データにベイジアンネットワークを適用した用途と技術の関係分析の事例

本節では、特許文書データに Nomolytics を適用した分析事例について、特に前節で述べた特許の用途と技術の関係を分析する事例について紹介します。なお、ここで紹介する事例は筆者のコンサルティング業務で実際にクライアントから依頼のあった分析の事例ではなく、あくまでも対外紹介用に作成した分析事例となりますが、ここで紹介している分析のアプローチは実際に筆者のコンサルティング業務でよく提供しているものです。

6.6.1 分析の趣旨

本分析では、国内の特許公報の要約文の情報を対象に、そこに記載されている課題の項目の文章と解決手段の項目の文章から、テキストマイニングと PLSA を適用することで、それぞれ用途に関するトピックと技術に関するトピックを抽出し、そのトピックを軸に特許データ全体の傾向を把握するとともに、用途トピックと技術トピックの関係をベイジアンネットワークでモデル化します。分析のプロセスの概要を**図 6.19** に示します。ここでは

（1）トピックの抽出
（2）トピックのスコアリング
（3）競合他社の分析
（4）用途と技術の関係分析

という四つのステップで特許文書データを分析します。それぞれの概要は以下の通りです。

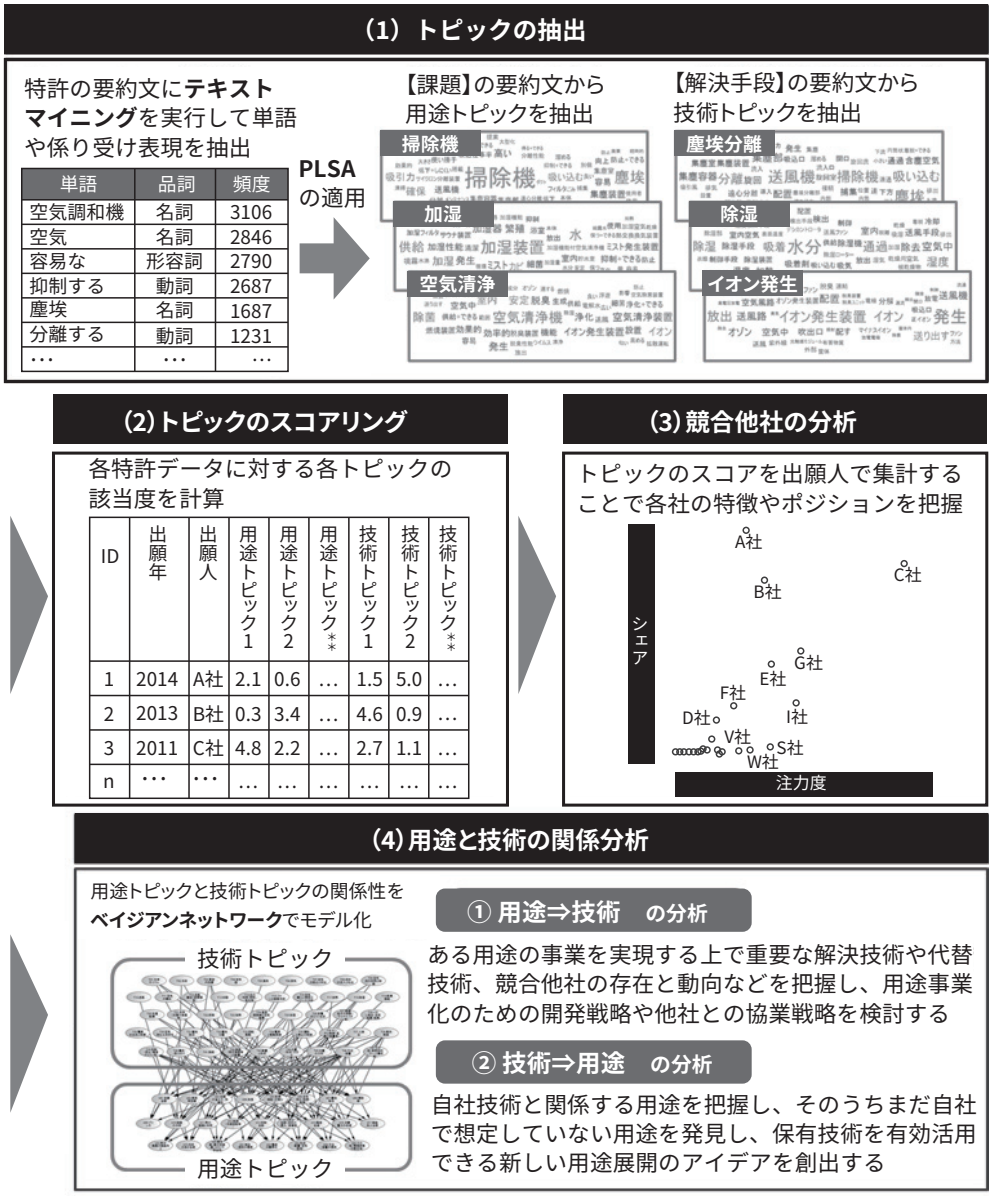


図 6.19　特許文書データに Nomolytics を適用した分析のプロセス

(1) トピックの抽出

特許の要約文に記述されている「課題」と「解決手段」という項目の文章を対象に、テキストマイニングと PLSA を適用し、それぞれ用途に関するトピックと技術に関するトピックを抽出します。ここで得られた結果により、分析対象の特許に記されている用途と技術の全体像を把握できます。

(2) トピックのスコアリング

分析対象としている各特許データに対して抽出したトピックのスコア（該当度）を計算します。ここで得られた結果により、各特許に対して抽出したトピックを軸に分類･整理できます。

(3) 競合他社の分析

トピックのスコアデータを「出願人」の軸で集計することで、各出願人のポジショニングや出願動向を分析します。ここで得られた結果により、自社の技術開発戦略や差別化戦略、他社との協業戦略、自社技術の売却先などを検討できます。

(4) 用途と技術の関係分析

用途のトピックと技術のトピックの確率的な因果関係をベイジアンネットワークでモデル化し、用途と技術の関係構造を分析します。ここでの分析は、①用途⇒技術の分析（用途に対して関係する技術の分析）と、②技術⇒用途（技術に対して関係する用途の分析）の二つのパターンがあります。

①用途⇒技術の分析は、自社で検討しているある用途の事業を実現する際、重要となる要素技術を把握するための分析です。ここで得られた結果により、その用途を達成するためにどのような技術開発に注力すべきか、またその技術領域で競合となりそうな他社はどこか、そのなかで他社が牛耳る技術の代替技術は存在するか、どの出願人と技術提携すると効果的かなど、自社の開発戦略や他社との協業戦略を検討できます。

②技術⇒用途の分析は、自社で保有している技術と関係のある用途を把握するための分析です。ここで得られた結果により、自社技術と関係のある用途のうち自社でまだ想定していない用途を見つけ、自社の技術をさらに有効活用できる新しい用途展開のアイデアを創出できます。

6.6.2 分析で用いるデータ

本節で紹介する事例では、要約と請求項に「風」「空気」という二つのキーワードを含む国内の特許公報データ 30039 件を分析対象としています。出願期間は 2006 年 1 月 1 日から 2015 年 12 月 31 日までのちょうど 10 年分の特許公報を対象に抽出しました。特許は出願し

てから公開されるまで 1 年半の期間を要するので、それを考慮して、データの抽出は 2017 年 7 月 25 日に実施したものです。

特許の要約文は 400 字以内という文字数制限があり、また厳密なルールではありませんが、【課題】や【解決手段】などの見出しをつけてから、それに該当する内容を記述することが慣例となっています。要約文の例を**図 6.20** に示します。「風」「空気」をキーワードに含む特許のデータなので、たとえばエアコンや扇風機、空気清浄機、加湿器、掃除機、洗濯乾燥機などさまざまな生活家電が関連した特許が含まれます。

【要約】【課題】ユーザーの快適性を維持しつつ、省エネ運転を行うことができる空気調和機を提供すること。【解決手段】本発明の空気調和機は、室内温度を検出する室内温度検出手段と、人体の活動量を検出する人体検出手段と、基準室内設定温度を設定するリモコン装置を備え、室内温度が基準室内設定温度となるように空調制御を行う空気調和機であって、人体検出手段で検出する活動量が所定の活動量以内であるときは、室内温度が、基準室内設定温度を補正した補正室内設定温度となるように空調を行い、補正室内設定温度よりも低い状態を継続すると、圧縮機を停止させ、圧縮機の復帰は、基準室内設定温度に基づいて行う。

※例示のための要約文であり、一部内容は筆者が加工しています。

図 6.20　分析対象とする特許の要約文の例

6.6.3　用途と技術のトピックの抽出

分析ではまずテキストマイニングと PLSA を用いて特許の要約文の内容をいくつかのトピックに集約します。先述の通り、国内特許の要約文では、【課題】と【解決手段】という二つの項目が記載されていることが多いため、課題の項目の文章と解決手段の項目の文章をそれぞれ抽出し、課題からは用途に関するトピックを、解決手段からは技術に関するトピックを抽出します。なお、課題と解決手段という項目のない特許は、要約文全体を使用します。

トピック抽出の手順をまとめると以下のようになります。

(1) テキストマイニング

課題の項目で記述された文章と解決手段の項目で記述された文章を切り出し、それぞれにテキストマイニングを適用し、単語と係り受け表現を抽出します。単語は名詞、動詞、形容詞、形容動詞を、係り受けは名詞に対する動詞・形容詞・形容動詞の文法的なつながりのある単語ペアを抽出します。なおテキストマイニングの実行には Text Mining Studio（株式会社 NTT データ数理システム）[2] を使用しています。

（2）共起行列の作成

続いて PLSA でトピックを抽出する際のインプットとする共起行列を作成します。6.4 節でも述べた通り、本来の PLSA では、文章（行）×単語（列）という構成の共起行列を用いますが、本事例で適用する Nomolytics では、単語（行）×係り受け（列）という構成で、単語と係り受けが同時に出現する共起頻度を集計した共起行列を用います。なお、共起行列の構成に採用する単語と係り受けは出現頻度 10 件以上を対象とし、「課題」の文章からは単語（3256 語）×係り受け（2084 表現）の共起行列を、「解決手段」の文章からは単語（5187 語）×係り受け（7174 表現）の共起行列を作成しました。なお、多くの特許で共通して使用される単語で、かつ頻度が高すぎる単語（「備える」や「提供する」など）や、重要な意味をもたない単語（「前記」や「本発明」など）は、トピック抽出においてノイズになり得るため、対象から除外しています。

（3）PLSA の実行

作成した共起行列に PLSA を適用することで、使われ方の似ている単語と係り受けでまとめられたトピックを抽出します。「課題」の共起行列からは用途に関するトピックを、「解決手段」の共起行列からは技術に関するトピックを抽出します。なお PLSA はあらかじめトピック数を設定する必要があります。また、最初に初期値も与える必要がありますが、その値により解が異なる初期値依存性があります。そこでトピック数を 1 刻みで変化させ、それぞれのトピック数に対して初期値をランダムに変えて PLSA を 5 回ずつ実行し、それぞれの解を情報量規準 AIC で評価して、最も評価の良い解を採用することにします。なお PLSA の実行には Visual Mining Studio（株式会社 NTT データ数理システム）の二項ソフトクラスタリングという PLSA を拡張させた同様の分析機能を使用しています。Visual Mining Studio は 6.4 節で紹介した同社製品の Alkano[13] がリリースされる前に販売されていたデータマイニングツールで、現在は販売を終了しており、後継製品が Alkano です。

6.6.4　抽出されたトピックの結果

上記の方法によりトピックを抽出した結果、用途については 25 個のトピックが、技術については 47 個のトピックが得られました。なお PLSA のアウトプットは、①各トピックにおける行要素（単語）の所属確率、②各トピックにおける列要素（係り受け）の所属確率、③各トピックの存在確率、という三つの確率が計算されます。抽出された用途と技術のトピックの内容の例を**表 6.2** に示します。単語と係り受けは所属確率の高い順に並べています。表 6.2 左の用途トピック U04 では、単語は、加湿装置、水、供給、加湿、カビなどが、係り受けは、加

湿装置⇒提供、加湿器⇒提供、ミスト発生装置⇒提供、水⇒供給、細菌⇒繁殖といった表現で所属確率が高いので、この結果は加湿に関するトピックであると解釈できます。また表 6.2 右の技術トピック T32 では、単語は、送風機、塵埃、掃除機、分離、吸い込む、集塵部などが、係り受けは、塵埃⇒分離、分離⇒塵埃、塵埃⇒含む、吸い込む⇒塵埃、含む⇒空気、空気⇒分離といった表現で所属確率が高いので、この結果は塵埃の分離に関するトピックであると解釈できます。このように解釈をつけた 25 個の用途トピックと 47 個の技術トピックの一覧をそれぞれ**表 6.3**、**表 6.4** に示します。

表 6.2　抽出されたトピックの例

用途トピックU04			
確率	単語	確率	係り受け
5.5%	加湿装置	6.8%	加湿装置⇒提供
3.7%	水	3.1%	加湿器⇒提供
3.3%	供給	2.9%	ミスト発生装置⇒提供
2.4%	加湿	1.9%	水⇒供給
2.3%	カビ	1.7%	細菌⇒繁殖
2.1%	加湿器	1.5%	加湿⇒行う
…	…	…	…

技術トピックT32			
確率	単語	確率	係り受け
5.5%	送風機	2.1%	塵埃⇒分離
5.2%	塵埃	1.7%	分離⇒塵埃
4.1%	掃除機	1.7%	塵埃⇒含む
3.6%	分離	1.5%	吸い込む⇒塵埃
3.5%	吸い込む	1.3%	含む⇒空気
2.3%	集塵部	1.0%	空気⇒分離
…	…	…	…

表 6.3　用途トピック一覧

No.	トピック名	No.	トピック名
U01	空調全般	U14	防止全般（流体の侵入、破損等）
U02	車両用空調	U15	騒音低減
U03	空調の省エネ、快適性	U16	消費電力の低減
U04	加湿	U17	機能向上全般
U05	乾燥機能（衣類等）	U18	熱交換器の機能向上
U06	空気浄化（除菌・脱臭）	U19	効率の良さ全般
U07	塵埃除去	U20	高性能・高付加価値（コストや安全性等）
U08	掃除機	U21	検出・測定の精度
U09	プリンタ	U22	構造の簡素化
U10	機器の冷却	U23	形成・配置（空気路等）
U11	熱の制御と利用	U24	方法・装置の提供
U12	制御（冷媒回路等）	U25	その他（環境破壊の懸念等）
U13	抑制全般		

表 6.4　技術トピック一覧

No.	トピック名	No.	トピック名
T01	冷凍サイクル	T25	加湿
T02	冷却	T26	放電式ミスト生成
T03	車室内空調	T27	微細粒子の飛散（マイナスイオン等）
T04	空気路	T28	イオン発生・空気除菌・脱臭
T05	換気	T29	電解水生成と除菌
T06	排気	T30	空気浄化&効率性
T07	空気の吸込と吹出	T31	塵埃除去
T08	流体の流入と吐出	T32	塵埃分離
T09	空気流の利用と制御	T33	回転駆動
T10	送風	T34	電源と駆動制御
T11	空気の噴出	T35	運転と停止の制御
T12	送風搬送（紙葉類等）	T36	センサと制御（温度や風量等）
T13	印刷	T37	人検出
T14	光の利用（照射、発光等）	T38	風向制御
T15	ファンと機器冷却	T39	抑制・防止（騒音やコスト等）
T16	空気導入と車両エンジンの冷却	T40	構成・取り付け
T17	放熱	T41	接続
T18	除湿	T42	機器（熱交換器等）の配置
T19	乾燥機能	T43	配置と形成
T20	洗濯乾燥	T44	位置・形状・大きさ
T21	洗浄（衣類や食器等）	T45	位置の方向
T22	燃焼	T46	方法・装置
T23	加熱	T47	その他（発明目的、ケース構成等）
T24	温湿度制御と空気循環		

6.6.5　トピックのスコアリング

続いて分析対象とした約 3 万件の特許データに対して、今回抽出された 25 個の用途トピックと 47 個の技術トピックのスコア（該当度）を計算します。スコアの計算では、1 件の特許の要約文には複数の文が存在するため、まず文単位（句点「。」で区切られた一文単位）に各トピックのスコアを計算し、それを特許単位に集約します。なお、文 S におけるトピック T のスコアは $P(S \mid T)/P(S)$ で定義します。これは事後確率と事前確率の比率を示し、リフト値と呼ばれることもある指標です。トピック T を条件とすることで、その文 S の発生確率が何倍になるのかを示すため、そのトピックをよく話題にしている文ほど値が高くなります。以下に $P(S \mid T)$ と $P(S)$ の計算方法について説明します。

$P(S \mid T)$ については、文 S を単語 X で定義される文 S_X と係り受け Y で定義される

文 S_Y に分解し、それぞれについて $P(S_X \mid T)$ と $P(S_Y \mid T)$ を計算し、それらを一つに統合して $P(S \mid T)$ を計算します。$P(S_X \mid T)$ と $P(S_Y \mid T)$ はそれぞれ式 (6.7) と式 (6.8) で計算されます。単語 X と係り受け Y が含まれる文の数をそれぞれ $N(X)$ と $N(Y)$ とすると、式 (6.7) の $P(S_X \mid X)$ は $N(X)$ の逆数、式 (6.8) の $P(S_Y \mid Y)$ は $N(Y)$ の逆数として計算されます。式 (6.7) の $P(X \mid T)$ と式 (6.8) の $P(Y \mid T)$ はそれぞれ PLSA の実行結果によって得られている単語と係り受けの所属確率に該当します。そして $P(S \mid T)$ は式 (6.9) で計算され、$P(S \mid S_X)$ と $P(S \mid S_Y)$ は文 S において重みは同じであるため、それぞれ 0.5 とします。また $P(S)$ は式 (6.10) で計算され、$P(T)$ は PLSA の実行結果によって得られているトピックの存在確率に該当します。

$$P(S_X \mid T) = \sum_X P(S_X \mid X)P(X \mid T) \tag{6.7}$$

$$P(S_Y \mid T) = \sum_Y P(S_Y \mid Y)P(Y \mid T) \tag{6.8}$$

$$P(S \mid T) = P(S \mid S_X)(S_X \mid T) + P(S \mid S_Y)P(S_Y \mid T) \tag{6.9}$$

$$P(S) = \sum_T P(S \mid T)P(T) \tag{6.10}$$

以上から $P(S \mid T)/P(S)$ で定義されるスコアを文単位に計算し、それを特許単位に見たとき、各トピックのスコアの最大値をその特許のトピックスコアとして採用します。さらにこのスコアの閾値を 3 に設定し、各特許データに対してそのトピックの該当有無を示す 1 と 0 のフラグ情報を付与します。なお、$P(S \mid T)/P(S)$ で定義したスコアは本来 1 が基準の目安になります。つまりスコアが 1 より大きいということはトピック T を条件にすることで文 S の確率が上昇するということで、トピック T と文 S の間に関係があると判定できます。本事例では各トピックの特徴を抽出するため、特に関連の強い特許に対して該当ありのフラグを立てることを考え、またこのスコアの分布や実際の文章の内容も確認しながら、その閾値は基準の 3 倍と厳しく設定しています。

以上の計算処理により、**図 6.21** に示すようなデータが作成されました。30039 件の特許データには、出願年、出願人、要約文という情報がもともとありますが、そこに加え、用途トピック 25 個、技術トピック 47 個の 1 と 0 のフラグ情報が付加されたデータとなります。このデータセットを用いることでトピックを軸にした、さまざまな分析を実行できます。

特許ID	出願番号	出願年	出願人	要約文		用途トピックU01	用途トピックU02	…	用途トピックU25	技術トピックT01	技術トピックT02	…	技術トピックT47
				【課題】	【解決手段】								
1	特願2006-XX	2006	A社	空気調和機の	吸気口から導入	1	0		0	0	1		0
2	特願2009-XX	2009	B社	短時間で除霜	着霜検出手段が	0	1		0	1	0		0
3	特願2011-XX	2011	C社	乾燥運転が可	通風路を通して	0	0		1	1	0		0
4	特願2013-XX	2013	D社	ウインドシー	車両用空調装置	0	1		0	0	1		1
…	…	…	…	…	…	…	…		…	…	…		…
30039	特願2012-XX	2012	Z社	プリ空調時に	冷暖房空調ユニ	0	1		0	1	1		0

図 6.21　トピックのスコア（フラグ情報）を紐づけた特許データ

6.6.6　トピックを用いた競合他社の分析

図 6.21 のトピックのスコアデータ（フラグデータ）を用いた分析の一例として、競合他社の分析を紹介します。ここでは出願人の情報と、トピックのフラグ情報から、各トピックにおける出願人の特徴を分析します。ここで得られた結果により自社の技術開発戦略や差別化戦略、他社との協業戦略、自社技術の売却先などを検討できます。

本分析では、各トピックにおいて出願人のポジショニングを可視化します。具体的には出願人 A とトピック T の関連度を示す「シェア」と「注力度」という二つの指標を計算し、縦軸にシェア、横軸に注力度を設定し、トピックごとに出願人をプロットしたポジショニングマップを作成します。シェアとは $P(A \mid T)$ で定義され、そのトピック T が該当する全特許の中での出願人 A の出願割合を示します。出願件数が多いほど値が高く、そのトピック T におけるシェアが高いということを意味します。注力度とは $P(T \mid A)$ で定義され、その出願人 A が出願した特許の中におけるトピック T の該当割合を示します。この値が高ければそれだけそのトピック T に注力しているということであり、独自の特有な技術を保有している可能性もあります。

ここでは技術トピック「T32. 塵埃分離」を例とした結果を**図 6.22** に示します。塵埃分離の技術とはサイクロン掃除機に代表される遠心分離などの技術になります。図 6.22 より、まずシェアで見ると、A 社、B 社、C 社が高いですが、特に C 社は注力度がとても高いため、特有の高い技術力を保有している可能性があります。一方、E 社、G 社、I 社などはシェアは中程度ですが、注力度が比較的高いため、技術力もある企業だと考えられます。この結果から、たとえば高いシェアの企業は、中程度のシェアの企業と連携することで、技術力を高めながらよりシェアを伸ばすことが期待できます。あるいは、中程度のシェアの企業の間で連携することで、業界大手に対抗するという戦略も考えられます。このように塵埃分離に関する技術は、三社のシェアが高いものの、他にもある程度のシェア・注力度を保有する企業が何社か存在し、今後の企業連携などの動きも考えられる領域と考察できます。

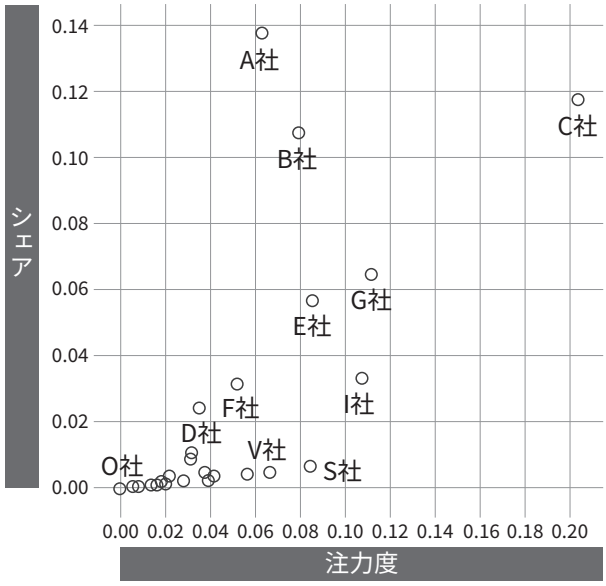


図 6.22　技術トピック「T.32 塵埃分離」における出願人のポジショニングマップ

6.6.7　ベイジアンネットワークを適用した用途と技術の関係分析

ここからはベイジアンネットワークを適用した用途と技術の関係分析について紹介します。図 6.21 のトピックのスコアデータを用いて、各用途トピックのフラグ情報と、各技術トピックのフラグ情報を確率変数とし、ベイジアンネットワークを適用することで用途トピックと技術トピックの関係構造を分析します。

本分析では、①用途⇒技術の分析（用途に対して関係する技術の分析）と、②技術⇒用途（技術に対して関係する用途の分析）という二つのパターンがあり、それぞれのベイジアンネットワークのリンク構造は逆転することになります。

①用途⇒技術の分析では、自社で検討しているある用途の事業を実現する際に、重要となる要素技術を把握するための分析です。ここで得られた結果により、その用途を達成するためにどのような技術開発に注力すべきか、またその技術領域で競合となりそうな他社はどこか、そのなかで他社が牛耳る技術の代替技術は存在するか、どの出願人と技術提携すると効果的かなど、自社の開発戦略や他社との協業戦略を検討できます。一方、②技術⇒用途の分析では、自社で保有している技術と関係のある用途を把握するための分析です。ここで得られた結果により、自社技術と関係のある用途のうち自社でまだ想定していない用途を見つけ、自社の技術を

さらに有効活用できる新しい用途展開のアイデアを創出できます。以下にそれぞれのパターンの分析結果と考察の例を紹介します。

6.6.8　用途⇒技術の関係分析

用途に対して関係する技術の分析では、用途トピック 25 個を親ノード（リンク元）に、技術トピック 47 個を子ノード(リンク先)に指定してベイジアンネットワークのモデルを構築し、用途に対する技術の関係構造を可視化します。**図 6.23** に構築されたモデルの結果を示します。

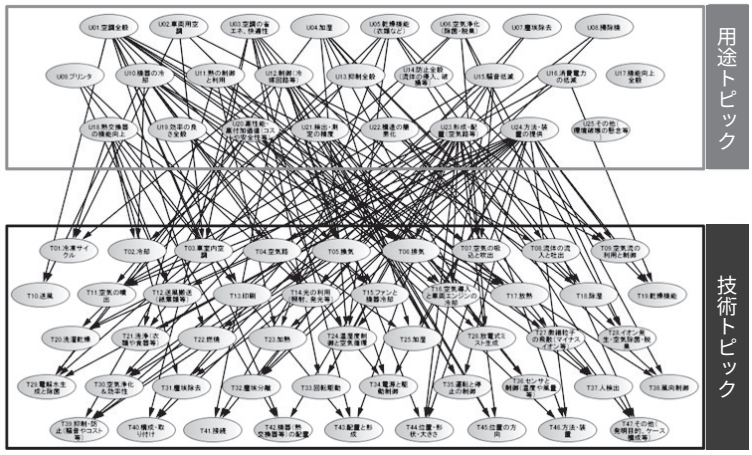


図 6.23　ベイジアンネットワークを適用した用途⇒技術の関係モデル

図 6.23 のモデルを用いることで、各用途トピックと確率統計的に関係があると判定された技術トピックを把握できます。またベイジアンネットワークの確率推論の機能により、ある用途トピックを条件に与えたときの各技術トピックの確率分布を推論できますが、特にその用途条件下で確率が上昇するような関係性の強い技術トピックを把握できます。ここでは用途トピック「U06. 空気浄化（除菌・脱臭)」を対象に関係の強い技術トピックを確認した例を紹介します。

図 6.23 のモデルを用いて、用途トピック「U06. 空気浄化（除菌・脱臭)」を条件に与えたときに、これと関係構造を持つ技術トピックの確率を推論した結果を**図 6.24** に示します。図 6.24 では、用途トピック U06 と関係が見られた各技術トピックのもともとの確率（事前確率）と、用途トピック U06 を条件に与えたときの条件付確率を掲載しており、どれも条件付き確率の方が高くなっているため、用途トピック U06 との関係が強いことがわかります。したがっ

て、「U06. 空気浄化（除菌・脱臭）」の用途と関係の強い技術トピックは、「T26. 放電式ミスト生成」、「T28. イオン発生・空気除菌・脱臭」、「T29. 電解水生成と除菌」、「T30. 塵埃吸込＆効率性」、「T47. その他」と確認できます。

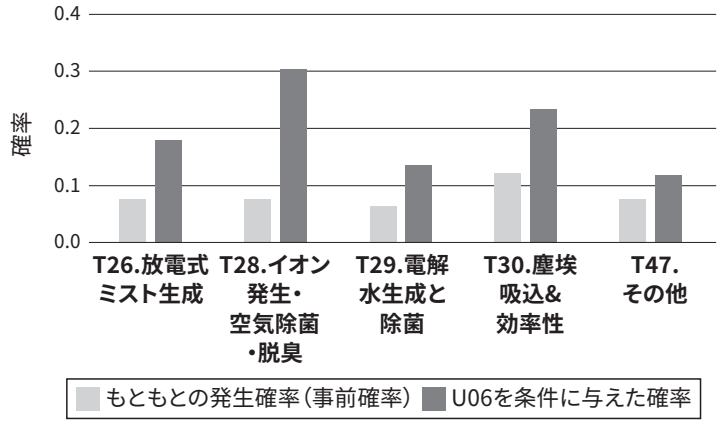


図 6.24　用途トピック「U06. 空気浄化」を条件に与えたときの各技術トピックの確率推論値

続いて、用途トピック「U06. 空気浄化（除菌・脱臭）」と関係が見られた「T.47 その他」を除く四つの技術トピックについて、各技術を保有している出願人を確認するため、前節で紹介した競合他社の分析を実施した例を解説します。ここでは、用途トピック「U06. 空気浄化（除菌・脱臭）」が該当する特許データを対象に、「T26. 放電式ミスト生成」、「T28. イオン発生・空気除菌・脱臭」、「T29. 電解水生成と除菌」、「T30. 塵埃吸込＆効率性」について、それぞれ各出願人のシェアと注力度を計算してポジショニングマップを作成しました。その結果を**図 6.25** に示します。

図6.25より、たとえば「T26. 放電式ミスト生成」は、シェアはA社とG社が高いですが、高シェア高注力度のポジションは空いていることがわかります。「T28. イオン発生・空気除菌・脱臭」と「T29. 電解水生成と除菌」は一社が高シェア高注力度のポジションを確立した一強状態にある技術領域であり、T28はG社が、T29はI社が牛耳っている技術であることがわかります。「T30. 塵埃吸込＆効率性」は、シェアはA社が高いですが、T26と同じく高シェア高注力度のポジションは空いています。

この結果より、たとえば一強状態の技術を避けて「U06. 空気浄化（除菌・脱臭）」の用途を実現するのであれば、T26やT30の技術が狙い目と考えることができるかもしれません。あるいは逆に一強状態にあるT28やT29の技術においては、その一強企業と提携する、あるいはM&Aを実現すれば、その技術領域ごと獲得できることになります。実際に「T29. 電解

水生成と除菌」を牛耳るI社はすでに買収されており、その買収した会社からは電解水（次亜塩素酸）で空気を洗うという新しい空気浄化家電が発売されていますが、その技術はI社で培われた技術であると考察できます。

このように自社で事業化を検討している用途に関係する重要な要素技術を分析することで、その用途を達成するためにどのような技術開発に注力すべきか、また競合となりそうな他社はどこか、他社が牛耳る技術を回避するような代替技術は存在するか、あるいはどの出願人と連携すると効率的にその技術を獲得できるかといった、開発戦略や協業戦略を検討できます。

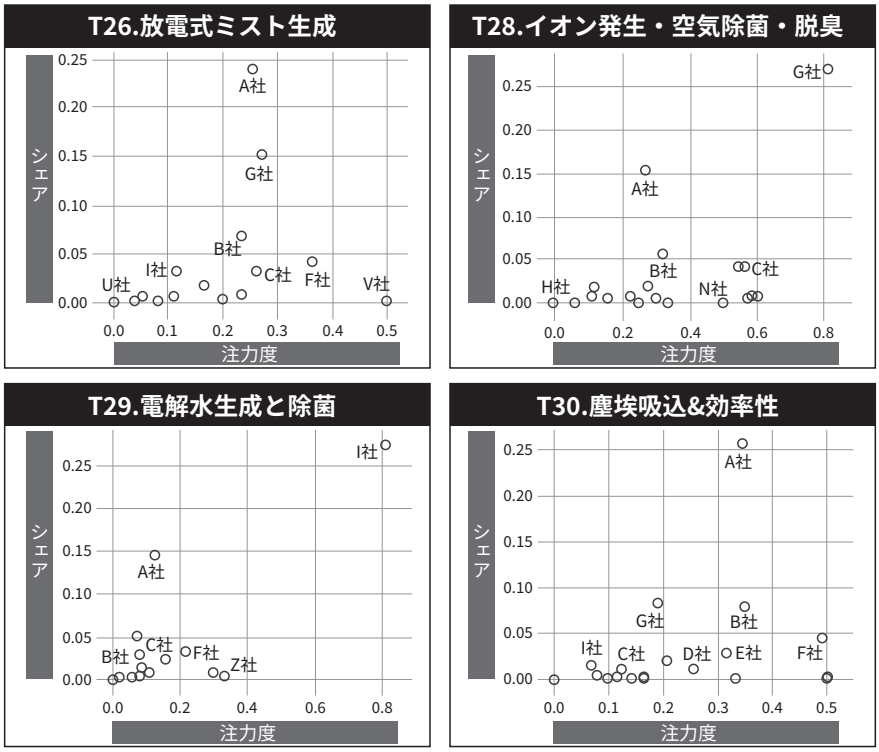


図 6.25　用途トピック「U06. 空気浄化」と関係のある技術トピックの出願人ポジショニングマップ

6.6.9　技術⇒用途の関係分析

技術に対して関係する用途の分析では、先ほどの用途⇒技術の関係分析におけるモデルのリンク構造を逆転させ、技術トピック 47 個を親ノード（リンク元）に、用途トピック 25 個を子ノード（リンク先）に指定してベイジアンネットワークのモデルを構築し、技術に対する用途の関係構造を可視化します。**図 6.26** に構築されたモデルの結果を示します。

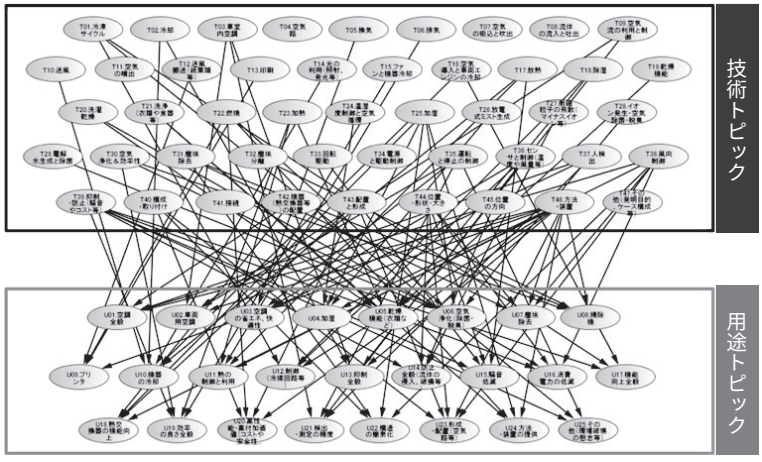


図 6.26　ベイジアンネットワークを適用した技術⇒用途の関係モデル

図 6.26 のモデルを用いることで、各技術トピックと確率統計的に関係があると判定された用途トピックを把握できます。また、ある技術トピックを条件に与えたときの各用途トピックの確率分布を推論できるため、特にその技術条件下で確率が上昇するような関係性の強い用途トピックを把握できます。ここでは技術トピック「T18. 除湿」を対象に関係の強い用途トピックを確認した例を紹介します。

図 6.26 のモデルを用いて、技術トピック「T18. 除湿」を条件に与えたときに、これと関係構造を持つ用途トピックの確率を推論した結果を**図 6.27** に示します。図 6.27 では、技術トピック T18 と関係が見られた各用途トピックのもともとの確率（事前確率）と、技術トピック T18 を条件に与えたときの条件付確率を掲載しており、どれも条件付き確率の方が高くなっているため、技術ピック T18 との関係が強いことがわかります。したがって、「T18. 除湿」の技術と関係の強い用途トピックは、「U05. 乾燥機能（衣類等）」、「U12. 制御（冷媒回路等）」と確認できます。

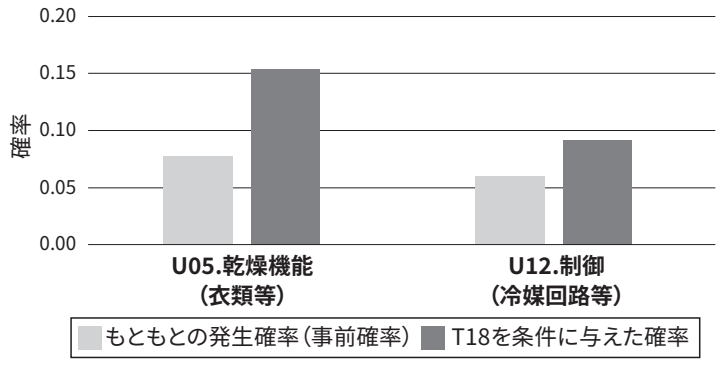


図 6.27　技術トピック「T18. 除湿」を条件に与えたときの各用途トピックの確率推論値

さらにここでは技術トピック「T18. 除湿」と用途トピック「U05. 乾燥機能（衣類等）」の関係に着目し、この技術の新しい用途展開を探索する考え方について紹介します。まず、この両者の関係の強さをデータ件数と割合で改めて確認すると、全体の特許（30039 件）の中で U05 の用途が該当するのは 2382 件あり、その該当割合 $P(\mathrm{U05})$は 7.9% となります。一方で、T18 の技術が該当する特許は全体で 2333 件あり、その中で U05 の用途が該当するのは 470 件あり、その条件付きの該当割合 $P(\mathrm{U05}\,|\,\mathrm{T18})$は 20.1% となります。つまり、T18 の技術を条件に与えることで U05 の用途の割合は 7.9% から 20.1% に上昇するので（その比率は $P(\mathrm{U05}\,|\,\mathrm{T18})/P(\mathrm{U05}) = 2.5$）、やはりこの技術トピック「T18. 除湿」と用途トピック「U05. 乾燥機能（衣類等）」は関係が強いことがわかります。

このことから、「T18. 除湿」の技術に関連する特許は、「U05. 乾燥機能（衣類等）」の用途を想定して出願されていることが相対的に多いということですが、逆にその傾向が成立していないケースを見つけることが技術の新規用途を探索するヒントになります。たとえば、ある出願人 X に注目すると、X 社は T18 の技術に該当する特許が 18 件ありましたが、このうち U05 の用途に該当するものはたった 1 件しかありませんでした。つまりベイジアンネットワークで明らかになった技術と用途の全体での関係性を見れば、この X 社の保有している「T18. 除湿」の技術はもっと「U05. 乾燥機能（衣類等）」の用途に展開できる可能性があると考察できます。

さらに実際の特許文書の内容を確認することで、この新規用途探索の分析をより深く、より具体的に進めることができます。まず「T18. 除湿」の技術が「U05. 乾燥機能（衣類等）」の用途を想定して出願されている特許の代表例を**図 6.28** 左に示します。これはドラム式洗濯乾燥機の特許ですが、特許の要約の内容を確認すると、洗濯物を短い時間でムラなく乾燥させ、乾燥工程の時間を短くするための除湿技術として出願されています。一方、出願人 X が出願している特許の中で「T18. 除湿」の技術に該当する特許の一例を取り上げたものを図 6.28 右に示

します。これはインクジェットプリンタに関する特許ですが、インク液を吸収した紙の湿気をムラなく取り除いて、コックリング（紙の波打ち）を防ぐ除湿の技術として出願されています。

実際にX社は洗濯乾燥機の製造はしていませんが、プリンタという空間の中でインク液を吸収した用紙の湿気をムラなく取り除いて紙の波打ちを防ぐ除湿技術は、たとえば洗濯乾燥機という空間の中で洗濯物をムラなく効率的に乾燥させることにも応用できる可能性を考えられます。

U05.乾燥機能を想定したT18.除湿の特許例	U05.乾燥機能を想定していないT18.除湿の特許例
【発明の名称】 ドラム式洗濯乾燥機 【課題】 洗濯物を短い時間でムラ無く乾燥させ、乾燥工程の時間を短くすることができるドラム式洗濯乾燥機を提供する。 【解決手段】 送風機に吸い込まれた空気は、風路切替弁の切り替えにより、ドラム開口部に対向する前側吹出口へ流れたり、回転ドラムの後部に設けられた後側吹出口へ流れたりする。制御装置が風路切替弁の切り替えを制御することによって恒率乾燥過程時、前側吹出口から乾燥用空気が吹き出し、かつ、減率乾燥過程時、後側吹出口から乾燥用空気が吹き出す。これにより、恒率乾燥過程において乾燥用空気が効果的に当たらなかった、回転ドラムの後端壁側の洗濯物に、乾燥用空気が減率乾燥過程で効果的に当たる。	【発明の名称】 インクジェット記録装置及び画像記録方法 【課題】 処理液の厚みムラを低減するとともに処理液による用紙のコックリングを低減することで、高品質かつ高速の画像記録を可能とするインクジェット記録装置及び画像記録方法を提供する。 【解決手段】 記録媒体に処理液を付与する処理液付与部の後段には、記録媒体表面に残存する溶媒を蒸発させるプレ加熱部が設けられている。プレ熱部はIRプレヒータにより記録媒体表面を輻射加熱するとともに、吸引ファンにより記録媒体表面の湿り空気を置換する。液状の処理液が不均一にならないように乾燥処理を施すことで、均一な膜厚を持つ個体状の擬集処理層が形成される。その後、本加熱部による熱風噴射加熱により、コックリング量が所定量以下になるように本加熱処理が施される。

※例示のための要約文であり、一部内容は筆者が加工しています。

図 6.28　技術「T18. 除湿」が異なる用途で出願されている二つの特許の例

もう一つ、技術に対する用途の関係に基づいた新規用途探索の分析例を紹介します。ここでは先ほどの競合他社分析でも取り上げた「T32. 塵埃分離」を対象に考察していきます。図 6.26 のモデルでは、「T32. 塵埃分離」の技術トピックと関係の強い用途トピックは、「U08. 掃除機」のただ一つでした。つまり、「T32. 塵埃分離」の技術に関する特許は、「U08. 掃除機」の用途を想定して出願されていることが相対的に多いということですが、ある出願人Yに注目すると、Y社はT32の技術に該当する特許が13件ありましたが、このうちU08の用途に該当するものは1件もありませんでした。つまり先ほどと同様に、全体における技術と用途の強い関係性を考えれば、このY社の保有している「T32. 塵埃分離」の技術はもっと「U08. 掃除機」の用途に展開できる可能性があると考察できます。

先ほどのように実際の特許文書の内容を確認してみると、まず「T32. 塵埃分離」の技術が「U08. 掃除機」の用途を想定して出願されている特許の代表例を**図 6.29** 左に示します。これはサイクロン掃除機の特許ですが、特許の要約の内容を確認すると、そのサイクロン掃除機の

排気筒の詰まりを防止して、集塵性能を向上させる技術として出願されています。一方、出願人Yが出願している特許の中で「T32. 塵埃分離」の技術に該当する特許の一例を取り上げたものを図 6.29 右に示します。これは画像形成装置、つまりプリンタに関する特許ですが、このプリンタではトナーが含まれる空気をサイクロンで分離して回収しており、そのサイクロン部の清掃時期をセンサで判断し、自動で清掃モードというものを実行することでトナーの分離効率の低下を抑制するという技術として出願されています。

U08.掃除機を想定したT32.塵埃分離の特許例	U08.掃除機を想定していないT32.塵埃分離の特許例
【発明の名称】 電気掃除機 【課題】 集塵性能が向上しメンテナンスの軽減が図れる電気掃除機を提供すること。 【解決手段】 塵埃を含む空気を旋回させ塵埃分離する略円筒状の1次旋回室と、1次旋回室に連通した2次旋回室と、1次旋回室の下方に位置し塵埃を溜める集塵室と、塵埃を圧縮する圧縮板と、塵埃が流入する流入口を有し、圧縮板の底面の一部に突出部を流入口から見て集塵室の奥側に配設する構成としたことより、集塵室内に入った塵埃は、圧縮板の突出部に引っかかり動きが止められ、流れに乗って2次旋回室や1次旋回室側に戻ることが無いため集塵性能が向上し、排気筒の詰まり防止によるメンテナンスの軽減を図ることができる。	【発明の名称】 画像形成装置 【課題】 サイクロン部の清掃時期を適正に判断して、トナーの分離効率の低下を抑制することが可能な画像形成装置を提供する。 【解決手段】 画像形成装置は、トナー含有空気からトナーを遠心分離するサイクロン部と、サイクロン部によって分離されたトナーを回収する回収部と、サイクロン部によってトナーが分離された空気を通過させ、残留トナーを捕集するフィルタ部と、空気を吸引する送風部と、フィルタの汚れを検知する汚れ検知センサが設けられたトナー捕集部を備え、汚れ検知センサで検知されたフィルタの汚れから推定した風量と、風速センサで取得した風量の実測値の差分が、サイクロン清掃閾値を超えたと判断すると、サイクロン部の清掃モードを実行する。

※例示のための要約文であり、一部内容は筆者が加工しています。

図 6.29　技術「T32. 塵埃分離」が異なる用途で出願されている二つの特許例

実際にY社はサイクロン掃除機の製造はしていませんが、プリンタの中でトナーを分離・回収するサイクロン部の清掃時期を判断して分離効率を維持する技術は、サイクロン掃除機の集塵部の集塵性能を向上させることにも応用できる可能性を考えることができます。

このように自社で保有している技術と関係のある用途を把握し、そのうちまだ自社で想定していない用途を見つけることで、自社の技術をさらに有効活用できる新しい用途展開のアイデアを創出できます。自社で注力してきた技術をそっくりそのまま別の用途に転用できるようなことはなかなかないかもしれませんが、これまで自社で培ってきた技術や経験と関連のある用途をいかに発想できるかということがイノベーションの鍵になります。ここで紹介した例はあくまでも分析結果から筆者が発想したアイデアであり、現実性は検討していませんが、こうした分析を実施していくことで、これまで発想していなかった新しい用途展開の気づきが得られるものと期待できます。

6.6.10　特許文書データの分析事例のまとめ

本節では、特許文書データに Nomolytics を適用した分析事例について、特にベイジアンネットワークを適用して特許の用途と技術の関係を分析する事例について紹介しました。ここでは Nomolytics を適用した特許文書分析の二つのメリットをまとめたいと思います。

まず一つ目のメリットは、PLSA を適用することで単語ではなく集約されたトピックを軸に分析を実行でき、膨大な特許情報に潜む傾向をわかりやすく理解できることです。従来のテキストマイニングのみを適用した特許文書データの分析では、そのアウトプットは図 6.1 に示したような、大量の単語をベースにした複雑なアウトプットとなるので、そこから傾向を把握することが難しいという課題があります。また、その大量の単語を人がグルーピングしていくつかのカテゴリを作成し、そのカテゴリをベースに分析することもありますが、そのカテゴリの作成が属人的で作業負荷が大きいという課題もあります。これに対して Nomolytics の分析では、特許文書全体に存在するトピックを PLSA で機械的に抽出して分類・整理できますし、単語ではなくトピックをベースに出願人の傾向などを分析すれば、その特徴をわかりやすく考察できます。

二つ目のメリットは、ベイジアンネットワークを適用することで、用途と技術の統計的な関係を把握でき、各用途を実現するための重要技術を確認して技術戦略を検討したり、自社技術を有効活用できる新規用途のアイデアを創出できることです。従来の特許分析でも、図 6.17 に示したように用途と技術の関係を分析するアプローチはありましたが、「課題」と「解決手段」それぞれに対して人がグルーピングして作成したカテゴリのクロス集計をし、その対応関係を考察するというもので、統計的な関係までは分析できていないという課題があります。これに対して Nomolytics の分析では、PLSA によって客観的に抽出されたトピックをベースに課題と解決手段の統計的な関係性をベイジアンネットワークで把握できます。そしてその構築された関係モデルを用いることで、たとえば、事業化を検討しているある用途に対して関係の強い技術を確認し、その技術の出願人の動向から自社の技術戦略を検討したり、あるいは自社の保有技術と関係の強い用途を見つけ、そこでまだ想定していない用途を確認することで、自社技術の新しい用途展開のアイデアを創出することなどに活用できます。

このように、従来のテキストマイニングに PLSA やベイジアンネットワークという二つの人工知能技術を組み合わせて特許文書データに適用することで、人間では読み切れない特許文書を複数のトピックに変換し、特許に潜む傾向をわかりやすく理解できます。また、用途と技術の統計的な関係を分析することで、用途に対する重要な解決技術の把握や技術に対する新しい用途を発想できるようになります。こうした分析結果を活用することで、企業の技術戦略の検討において新しい知見の創出が期待できます。

6.7 おわりに

本章では、テキストデータにベイジアンネットワークを適用して分析することの可能性と効果について、その実現アプローチと適用事例を紹介しながら解説しました。

テキストデータはあらゆる業界で活用の場面があり、それを分析することには大きな意義がありますが、本章で解説した通り、従来のテキストマイニングに加えてベイジアンネットワークを適用することでその有用性がさらに進化します。ベイジアンネットワークのモデリング手法としての特徴は、目的変数と説明変数の区別なく互いの変数の間にある関係性をネットワーク構造でモデル化でき、ある変数の状態を条件として与えたときの他の変数の起こり得る確率をさまざまな方向から推論できることにあります。テキストデータにこのベイジアンネットワークを適用することで、より具体的には、テキストマイニングで抽出された単語をトピックモデルで集約したデータにベイジアンネットワークを適用することで、テキストデータに潜む要因関係を構造的にモデル化でき、与えた条件に対する結果の挙動を確率的にシミュレーションできるようになります。これにより、たとえば実施する施策を条件として与えたときの効果を定量的に評価したり、その効果を最大化させる条件を探索でき、ビジネスの問題解決において効果的なアクションを検討する強力なツールになりえます。

このように、ビジネスの問題解決においてベイジアンネットワークの適用は有効に働くことが期待されますが、その使い手には認識すべきこと、注意すべきことがあります。最後にベイジアンネットワークを適用する使い手の心得として、①結果をミスリードしないための手法の理解、②問題解決におけるモデル化という行為の位置づけの理解、という 2 点の注意事項を述べることで本章の結びとしたいと思います。

6.7.1 心得 1：結果をミスリードしないための手法の理解

ベイジアンネットワークのモデル自体は、BayoLinkS を使用してデータをインプットすれば簡単に構築できます。しかし、その結果が本当にデータの真実を反映していると解釈するには注意が必要です。モデル自体は簡単に構築できるがゆえに、ベイジアンネットワークという手法の理論を理解していないと、間違った解釈や間違った使い方をしてしまうリスクもあります。特にモデルの結果は、データや変数の内容、設定条件、評価指標が少し異なるだけで大きく変化する可能性があります。ベイジアンネットワークの使い手はその可能性を心得て、なぜそのような結果が得られたのか、それは自分が意図して設定した条件で得られた結果なのか、

理解できていることが必要です。

たとえば、ベイジアンネットワークで用いられるカテゴリ変数は、 そのカテゴリの分け方でモデルの結果は異なります。ここで注意すべきことは、カテゴリ数の少ない変数ほど親ノードに選ばれやすくなることがあるということです。複雑度に応じたペナルティ項を含む AIC などをモデル構築の評価指標に使用した場合、カテゴリ数が少ない変数ほど条件付き確率表がシンプルになってペナルティ項が小さくなり、相対的にその変数の親ノードとしての評価が高くなります。そのため性別などカテゴリ数が 2 の変数から大量のリンクが張られることがよくあり、この結果をそのまま受け取ると、この変数があらゆる変数に関係をもつ重要な変数だと思ってしまいます。確かに重要な変数であることもありますが、単に変数のカテゴリ数が少ないがために、モデル構築の評価指標で相対的に有利になっているだけということもあります。このミスリードを回避する方法として、モデルを構築するときは、変数のカテゴリ数をなるべく統一することが挙げられます。

また、BayoLinkS には欠損のあるデータを処理する機能があります。たとえば変数と変数のつながりを評価するときだけ、それに該当する欠損を削除して評価するペアワイズという処理機能があります。ペアワイズを適用すると、欠損が一つでもあるデータをすべて削除するリストワイズという方法を取らなくても、変数の探索のときに該当する部分だけ削除することになります。このことは一見すると欠損データを有効活用できてよいと思いますが、注意すべきことがあります。それは欠損がある変数の方が評価指標において有利になり、リンクが張られやすくなることがあるということです。尤度を含む評価指標の場合、欠損がある変数ペアの方が、尤度における確率を計算する観測データの対象が少なくなることで尤度が高くなり、相対的に評価指標において有利になることがあります。もちろん尤度は推定確率にも依存しているので、単純に観測データが少ないだけで評価が上がるとは一概に言えませんが、欠損データを含む場合のモデル構築ではこうした現象はよく起こります。本来であれば、欠損がある方が情報の信頼性が下がるはずですが、欠損がある変数の方がリンクが張られやすくなれば、まるで欠損がある変数が重要であるかのようにミスリードしてしまいます。

このミスリードを回避する方法として、インプットにするデータにはなるべく欠損は含めないようにすること、欠損を多く含む変数は学習の対象から除外することなどが挙げられます。また欠損を補完するという対策も考えられ、平均値や中央値や最頻値の代入といった方法が簡易的な欠損補完の方法としてよく用いられます。また一方で、こうした従来の代入法は分析結果の推定値にバイアスが生じるおそれがあることも指摘されています。近年ではそうしたバイアスを抑える欠損処理法として多重代入法という方法も提案されており、疫学の分野などではその検証が取り組まれています [18]。どうしても欠損を含むデータを使用しなければならないときは、こうした代入法も参考になるかと思います。

またベイジアンネットワークはこうした結果の解釈だけでなく、結果の使い方にも注意が必

要です。ベイジアンネットワークのリンクは有向リンクになるので、これが本当の因果関係の向きであると勘違いしてしまう人がいます。これはあくまでも確率的な因果関係であり、二つの変数の間の関係は、リンク元を条件にした方が、リンク先を条件にするよりも、その条件付き確率による尤度が高くなるというだけで、現象における原因と結果の関係ではないということです。ベイジアンネットワークではモデル構築の学習の際にリンクの向きも自動判断させられるので、この機能を使って因果関係を客観的に抽出できると誤った考え方をする人もいます。ベイジアンネットワークで判断されるリンクの向きは本当の因果関係ではありません。

もし因果関係の向きに意味があるデータであれば、その向きの判定は機械任せにせず、あらかじめその因果の向きの仮説を立て、それに応じた親ノード候補の変数と子ノード候補の変数を分けて指定することが有効です。6.6 節で紹介した特許の用途と技術の関係分析では、①用途⇒技術の分析と②技術⇒用途の分析があり、両者のモデルは用途と技術のリンクの向きを逆に設定しました。これはそれぞれで分析の目的が明確に異なり、その分析目的に対して用途と技術の因果関係の向きが重要な意味をもつため、あらかじめリンクの向きを指定してそれぞれモデルを構築しました。一方、変数の向きに意味をもたないときはそうした指定はせず、リンクの向きを機械に任せることもあります。そのように構築されたモデルを考察するときは、リンクの有無だけに着目して、リンクの向きは無視したほうがミスリードを回避できます。

このようにベイジアンネットワークは、データや変数や設定条件が少し異なるだけで結果が全く変わってしまうこともあり、モデル構築のコントロールが難しい手法です。しかしモデル自体は簡単に構築できてしまい、またその結果も視覚的にわかりやすいため、そこからインサイトを得たいというモチベーションも加わって、結果をミスリードしてしまうリスクがあるということを認識しておく必要があります。こうしたミスリードを回避するため、ベイジアンネットワークの使い手は、モデルがどのような仕組みで構築されているのかというアルゴリズムの数学を少しでも理解し、なぜそのような結果になったのか、どのような条件にすれば獲得したいモデルを構築できるのかという、仮設を立てられるような分析スキルを身につけていることが求められます。

6.7.2　心得 2：問題解決におけるモデル化という行為の位置づけの理解

ベイジアンネットワークはビジネスの問題解決において有効に働くと先述しましたが、問題解決という取り組みのなかでベイジアンネットワークのモデル化とは何をしているのかということを理解していることが重要です。結論から話をすると、ベイジアンネットワークのモデル化は「問題の発見」をするための行為であり、「問題の解決」をしているわけではないということです。ベイジアンネットワークの使い手は、この違いを心得ておくことが肝要です。

ベイジアンネットワークを適用して得られるアウトプットは要因関係の構造モデルですが、モデルとはデータに記された現象に潜む特徴や傾向を「抽象化」したものです。抽象化をするということは、問題解決の取り組みにおいて「問題の発見」に役立ちます。問題を発見するということは、問題の本質を捉えることであり、それは個別事象ばかりに目を向けていても見えてくるものではありません。現象をメタ視点で抽象化して、そもそもどんな特徴や傾向が本質的に存在しているのかということを見極める必要があります。つまりモデルとは、現象の表面的な部分を並べているものではなく、その奥にある本質的な部分を映しているべきで、そうして構築されたモデルを用いることで問題の本質を捉えることができます。そのため、図 6.4 で示したような、とても複雑で構造を解釈できないようなモデルは、本来モデルとしては好ましくありません。

このように問題解決の取り組みにおいて、モデルとは「問題の発見」には役立ちますが、それだけでは「問題の解決」にはなっていません。問題を解決するアクションは必ず具体性を伴っているものであり、問題の解決を検討する際には、発見した問題の本質から解決策を「具体化」する必要があります。もともと具体的な個別事象のデータに対して、問題の本質を発見しやすくするためにモデル化によって「抽象化」しましたが、そこで捉えた問題から、具体的にどの問題をどうやって解決するのか、具体的にどんなアクションを実行するのかという、解決に向けて再度「具体化」を施すこと、つまり「具体」⇒「抽象」⇒「具体」というプロセスが求められます。モデルを構築した時点では、それは問題を発見するための抽象化された結果を得ただけであって、そこから問題を解決するには効果的なアクションを具体的に検討できなければいけません。データ分析というと、モデル化をして終了してしまう取り組みが多々あります。そうした結果は報告しても「で、だから何？」というリアクションが返ってしまいがちです。それはモデル化という問題の本質を「抽象化」することで分析の取り組みが終了してしまっており、そこから再度「具体化」した問題の解決につなげられていないからです。たとえば、6.6 節の最後で紹介した技術に対する用途の関係分析では、自社の技術がどんな用途と関係しているのかということをモデルで確認し、そこから関連する個別の特許データの中身に立ち返るこ

とで、自社技術の新規用途アイデアを具体的に発想することにつなげました。

そうした再度具体化するプロセスはまさに人間がやらなければいけないもので、そのプロセスには問題解決の思考スキルが求められます。問題解決の目的意識を強くもち、モデル化で確認できた問題の本質を前提に、問題解決のゴールに向かって、具体的にどのような方法がとれるのかという仮説を立てられる発想力が問われます。実はこのとき、ベイジアンネットワークはその問題解決の「具体化」の検討でもサポートしてくれる機能があります。それが確率推論によるシミュレーションです。6.5 節では、さまざまなモデル構築の例とモデルの確率的なシミュレーションによる活用例を紹介しました。たとえば温泉地のレビューデータの分析では、男女によって満足度を高める同行者や観光コンテンツに違いがあることがモデルからわかり、ターゲットに応じた具体的なプロモーション施策の例を紹介しました。構築されたモデルを用いてどのような条件でシミュレーションを実行するかは人間の判断が必要で、それには問題解決の目的意識の高さと具体的な仮説の発想力が求められます。一方、与えた条件の効果の評価や、高い効果を得られる条件の組み合わせは、ベイジアンネットワークのシミュレーション結果によって定量的に確認できます。つまり、モデル化によって問題の本質を理解し、そこから人間が問題解決の施策パターンを具体的に挙げることができれば、それぞれの施策の効果をモデルでシミュレーションし、効果的な施策を定量的に選択できます。

以上から、ベイジアンネットワークという手法は、モデル化によってデータで記された現象の本質部分を「抽象化」し、「問題の発見」に貢献できる一方で、そこから再度「具体化」して「問題の解決」を導くことはできません。それは人間の仕事であり、使い手の問題解決の思考スキルが求められます。ですが、ベイジアンネットワークはそんな人間が考える具体的な問題解決のアクションの効果性をシミュレーションすることもできるため、問題解決の補助となる有力なツールと捉えることができます。一方で、構築したモデルがそのまま問題解決につながるケースも存在します。たとえば深層学習を使って画像データの識別モデルを構築することは、そのアウトプットがそのまま画像の識別という問題の解決に直結しています。6.2 節で紹介した大規模言語モデルも、文章要約や言語翻訳など、そのアウトプットそのものが問題解決に直接的な価値を提供するものでした。こうしたケースの問題解決プロセスは、先ほどと対比させれば「具体」⇒「具体」と考えることができ、人間が問題の本質を理解して問題の解決を講じる必要はありません。実際に構築されたモデルはブラックボックスになることが多く、そもそも理解するためのものにはなっていません。このように、問題の本質の抽象化という問題発見ステップを踏まなくても、機械学習で得られた結果がそのまま問題解決になっているケースはとても扱いやすいといえます。しかし、ビジネスの問題解決では、本質を理解していないと解けない問題はたくさんあります。こうした問題解決にベイジアンネットワークは本領を発揮する手法だと考えることができます。

参考文献

[1] 那須川哲哉：『テキストマイニングを使う技術 / 作る技術―基礎技術と適用事例から導く本質と活用法』，東京電機大学出版局（2006）

[2] 株式会社 NTT データ数理システム："Text Mining Studio"，https://www.msi.co.jp/solution/tmstudio/index.html [2023 年 5 月 25 日閲覧]

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: "Attention is All You Need," Proc. of the 31st International Conference on Neural Information Processing Systems (NeurlPS), pp. 6000–6010 (2017)

[4] D. Bahdanau, K. Cho, and Y. Bengio: "Neural Machine Translation by Jointly Learning to Align and Translate," Proc. of the International Conference on Learning Representations (ICLR) (2015)

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186 (2019)

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever: "Improving Language Understanding by Generative Pre-Training," Open AI technical report (2018)

[7] 野守耕爾，北村光司，本村陽一，西田佳史，山中龍宏，小松原明哲："大規模傷害テキストデータに基づいた製品に対する行動と事故の関係モデルの構築－エビデンスベースド・リスクアセスメントの実現に向けて－"，人工知能学会論文誌，Vol. 25，No. 5，pp. 602–612（2010）

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman: "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391–407 (1990)

[9] T. Hofmann: "Probabilistic Latent Semantic Analysis," Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, pp. 289–296 (1999)

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan: "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, pp. 993–1022 (2003)

[11] 野守耕爾："テキストマイニングに複数の人工知能技術を応用した特許文書分析と技術戦略の検討"，情報の科学と技術，Vol. 68，No. 5，pp. 332–337（2018）

[12] 特許第 6085888 号："分析方法、分析装置及び分析プログラム"，（2017 年 2 月 10 日登録）

[13] 株式会社 NTT データ数理システム："Alkano"，https://www.msi.co.jp/solution/alkano/index.html/ [2023 年 5 月 25 日閲覧]

[14] 野守耕爾，神津友武：“三位一体アプローチによるテキストデータモデリング法の開発 ―宿泊施設の口コミデータを用いた評価推論モデルの構築―”，2014 年度人工知能学会全国大会論文集（2014）
[15] 野守耕爾，神津友武：“口コミビッグデータに人工知能を応用した地域観光の次世代マーケティング”，2016 年度人工知能学会全国大会論文集（2016）
[16] 新井喜美雄:“特許情報分析とパテントマップ”, 情報の科学と技術, Vol. 3, No. 1, pp. 16–21（2003）
[17] 安藤俊幸：“テキストマイニングと統計解析言語 R による特許情報の可視化”，情報の管理，Vol. 52，No. 1，pp. 20–31（2009）
[18] 早山陽子，山本健久，小林創太，村井清和，筒井俊之：“疫学調査データにおける欠損値が解析結果に与える影響の評価”，獣医疫学雑誌，Vol. 20，No. 2，pp. 111–117（2016）