第５節　特許文書のトピック分析と新たな技術戦略の切り口

はじめに

従来、特許分析というとパテントマップ[1]と呼ばれる手法が代表的であり、特許調査において長く親しまれている手法である。パテントマップは主に出願人や出願年、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、技術の実態や傾向を視覚的に容易に把握できるようにしたものである。また、特許の要約文や請求項、明細書といった文書情報も組み合わせてパテントマップを形成することもある。そこではテキストマイニングを適用し、人間ではなかなか整理することが難しい特許の文書情報から、より細かな特許技術の特徴を把握する。テキストマイニングは非構造化データであるテキストデータを統計的に分析可能な形にする自然言語処理技術であり、テキスト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ある[2]。その単語や係り受けの出現頻度を集計したり、出現関係をネットワーク図やマップ図で可視化することで、テキストデータの全体像の特徴を出現単語をベースに把握することができる。特許文書にテキストマイニングを適用することで、例えば、特許文書の記述内容の類似性に基づいたポジショニングマップを作成して各特許の位置づけを把握したり[3]、課題と解決手段に相当するキーワード群から分類軸を設定してそのクロス集計をすることで課題解決マトリクスを作成し、用途と技術の関係性を把握する分析などもある[4]。現在では様々な IT ベンダーからこうした分析ツールが販売されており、分析事例が多数報告されている[5]。特許の文書情報は人間が内容を読んで理解するだけでなく、テキストマイニングと統計解析を施してパテントマップとして可視化することは、特許分析の普遍的なアプローチになってきている。

さらに近年では特許文書データに AI を適用した分析の試みも増えてきている。昨今の第三次 AI ブームの中では、より性能の高い機械学習アルゴリズムが開発・公開され、目まぐるしい勢いで新たな AI 技術が誕生している。第三次 AI ブームの火付け役となったのはディープランニング[6]といえるが、最近では BERT[7]と呼ばれる自然言語処理技術も誕生し、文書分類や翻訳、検索などで高い精度の実績を上げている。特許文書分析にもこのような最先端の AI 技術を適用した様々な高度な分析アプローチが生まれている。そうしたアプローチに多く共通していることは、特許文書を一度ベクトル化して、そのベクトル化データを用いて教師あり学習の分類問題を解いたり、教師なし学習のクラスタリングを実行するという分析である。文書のベクトル化については、もっとも単純な手法として Bag-of-Words と呼ばれるものがあり、これは各文書においてある単語が登場する場合は“1”、登場しない場合は“0”というように、各単語の登場有無をフラグ化したデータによってベクトル化する手法である。このように文書をベクトル化することで定性的なテキス

トデータでも統計解析や機械学習を実行することが可能となる。自然言語処理の分野では、近年特にこの文書のベクトル化において様々な手法が提案されており、特許文書分析でも活用されている。例えば、doc2vec を用いて文書をベクトル化したデータを用いて特許分類を検討したり[8]、self-attention 機能を有する LSTM モデルで文書をベクトル化したデータで技術俯瞰マップを作成したり[9]、BERT により文書をベクトル化したデータで二つの請求項の同一性を判定する取り組みなどもある[10]。

こうした技術駆動型とも言える分析が活発化するなかで、従来のパテントマップのアプローチが陳腐化しているかというと、そうではない。そもそもパテントマップは分析者が母集団の特許に対して技術の全体像を把握するための手法であり、出願人や出願年、分類コード、キーワードなど、様々な軸から特許の傾向を可視化することで、関連技術の実態を人間が詳しく探索することを目的としている。こうしたパテントマップを用いることで、新しい技術の動向や未着手の技術の発見、競合企業との差異や関係性を把握でき、企業における技術戦略を人間が検討する上で有益な気づきを与えてくれる。そのため、従来のパテントマップは可視化の仕方こそ単純ではあるものの、それ故にとてもシンプルで人間が容易に理解しやすく、誰もがその分かりやすい可視化結果から着想を得ることができ、ビジネス活用面で優れたアプローチといえる。ビジネスにおいてデータを活用することの狙いの本質は課題解決にあり、データ分析は高度なスキルを有する限定された人間だけが扱えるものではなく、誰もが容易に理解でき、誰もが課題解決策について考察できるものが望まれる。リテラシーのハードルが高すぎるとビジネスにおけるデータ活用は進まない。企業が特許情報を分析することのゴールが有益な技術戦略の検討であることを考えれば、これまで長く親しまれてきたシンプルで分かりやすいパテントマップは大変有用といえる。

一方で、特許の文書情報を用いたパテントマップの従来の分析アプローチは、分析対象となる特許データの量が多くなると対応することが難しくなる。これまでの分析では人手を介した丁寧な処理もされるが、データ量が多くなるとこれを人間が一つ一つ実施することは現実的ではなく、分析結果に主観的な偏りも強く生じてきてしまう。また、テキストマイニングでは読み込むテキストの量が多くなれば、それだけ大量の単語が抽出されるため、その単語をベースに可視化をすればとても複雑な結果となってしまい、解釈の容易性が損なわれてしまう。

本稿では、複数の AI 技術を応用することで、大量の特許データであっても従来のパテントマップのように分かりやすいアウトプットを獲得できるような特許文書分析のアプローチとその適用事例を紹介する。具体的にはテキストマイニングに PLSA（確率的潜在意味解析）とベイジアンネットワークという 2 つの AI 技術を応用することで、企業の技術戦略の検討において新たな知見を創出する切り口を提供する。

1．特許の文書情報を用いた従来のパテントマップ分析

特許文書にテキストマイニングを応用した従来のパテントマップの分析において、その分析目的によく取り上げられるものとしては、(1) 全体像を把握する、(2) トレンドを把握する、(3) 競合他社の動向を把握する、(4) 用途と技術の関係を把握するといったものが挙げられる。アウトプットとしてよく用いられるものを**図 1** に示しながら、それぞれの分析の概要について以下に述べる。

(1)　全体像の把握

対象となる技術領域の全体像を俯瞰して把握するための分析である。最も基本的なものでは、**図 1** (A) のように特許文書に含まれる単語や文法的な単語のペアとなる係り受けの出現頻度を集計し、頻度の多い言葉から全体像を把握する。また**図 1** (B) のように各単語が同じ特許文書に出現する共起関係をネットワークで可視化し、単語のかたまり状況から、形成されている話題について定性的に考察する。

(2)　トレンドの把握

今後成長が見込まれる技術や、逆に衰退している技術を把握し、研究開発戦略を検討していくための分析である。抽出した単語の出現頻度を出願年で集計することもあるが、一つ一つの単語で集計すると複雑になりすぎるため、しばしば**図 1** (C) のように複数の単語や係り受けを人がグルーピングして意味性のあるカテゴリを形成し、**図 1** (D) のようにカテゴリ別に該当特許数を出願年で集計してそのトレンドを把握する。

(3)　競合他社の動向の把握

他社と差別化するような研究開発戦略を検討したり、自社技術のライセンス先候補や技術提携先候補となるような企業を検討するうえでニーズがある分析である。コレスポンデンス分析や数量化Ⅲ類と呼ばれる手法がよく用いられ、**図 1** (E) のように単語と出願人を同じ平面上にマッピングし、出願人と近くに位置する単語から各出願人の技術開発動向を把握する。

(4)　用途と技術の関係の把握

特に自社技術の新たな用途展開を探索するうえでニーズがある分析である。国内特許の要約文で記述されていることの多い「課題」と「解決手段」という 2 つの項目に着目し、それぞれの文章をテキストマイニングして**図 1** (C) のようにカテゴリを形成し、**図 1** (F) のように課題のカテゴリと解決手段のカテゴリに該当する特許件数をクロス集計することで、用途と技術の対応関係を把握する。

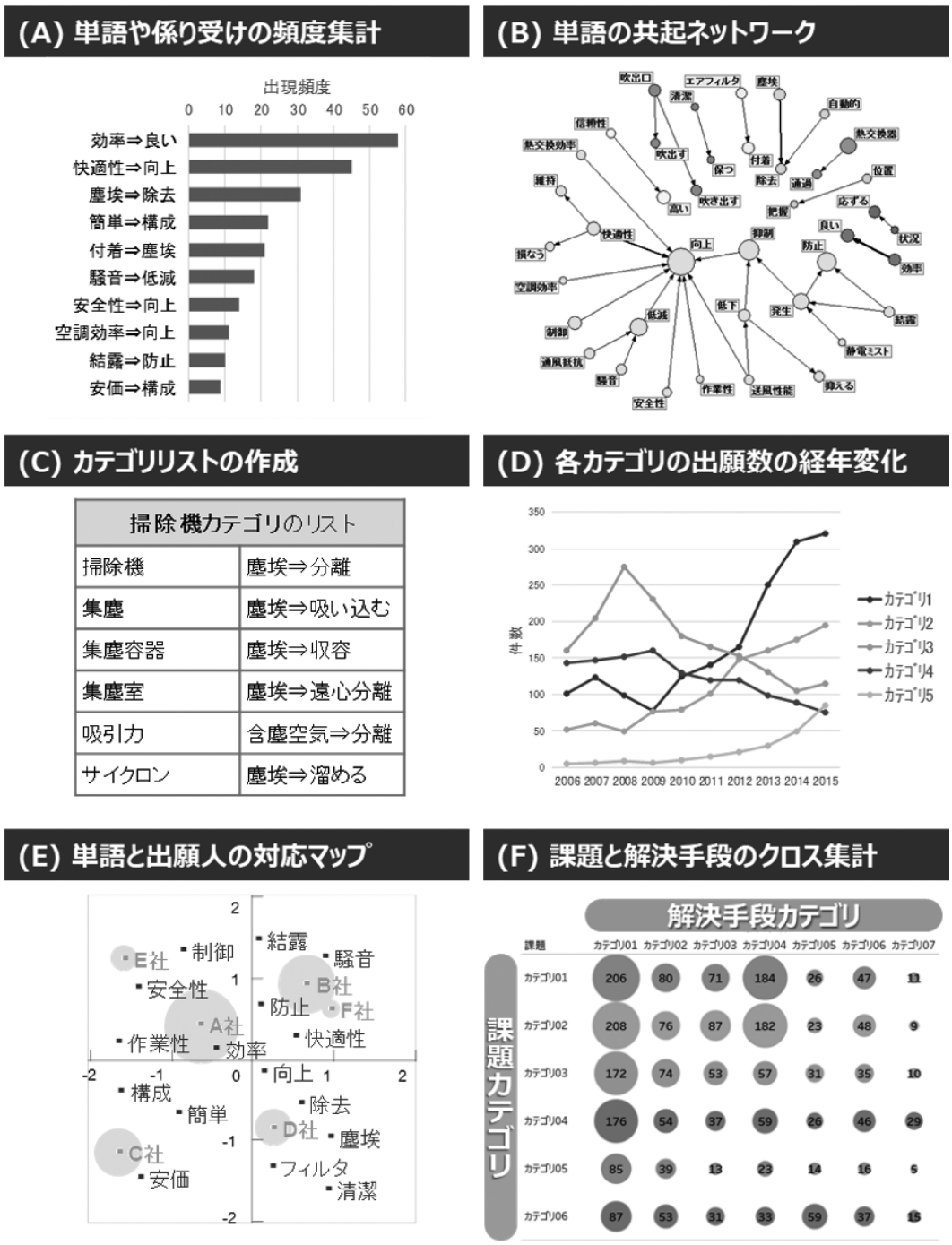


図１　従来の特許文書分析におけるパテントマップの可視化例

２．従来の特許文書分析の課題と AI 技術の応用

これらの分析は、人間ではなかなか読み切れない特許文書の全体像を把握するうえで有効な手段である一方、①基本的に単語をベースにした分析であるため単語数が増えると結果が複雑で考察しにくい、②単語のカテゴリの設定が主観的で作業負荷も大きい、③用途と技術の関係は単純なクロス集計で統計的な関係が分析できていない、といった課題もある。①の課題については、特にビッグデータとなるような件数がとても多い特許文書を分析する場合、テキストマイニングで抽出される単語も膨大となるため、大量の単語をベースにした可視化結果は複雑で解釈困難となることが多い。②の課題については、そのカテゴリルールが属人的となるため、作業者が変わればルールも変わってしまう曖昧なものであり、知識継承もされにくい。特に分析対象がビッグデータの場合、人間がその結果を整理してカテゴリルールを作成するにはあまりにも負荷が大きい。③の課題については、単純な頻度の大きさ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らない。つまりその用途や技術に該当する特許の元々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る。

こうした従来の特許文書分析の課題に対して AI 技術を応用することは一つの解決手段になることが期待できる。本節では、上記の①②の課題を解決する技術として、単語をクラスタリングできる PLSA（確率的潜在意味解析）という AI 技術を紹介する。PLSA では、単語群を集約したトピックを抽出することで、**図１(C)** のように従来人手で作成していた単語のカテゴリを機械的に実行できる。また、特許文書全体の傾向を従来のように大量の単語ベースに可視化をするのではなく、抽出されたいくつかのトピックをベースに可視化することで、シンプルで分かりやすい結果を得ることができる。こうしたトピックはパテントマップにおける新たな探索軸と捉えることもできる。一方、③の課題を解決する技術としては、変数間の要因関係を確率統計的にモデリングできるベイジアンネットワークという AI 技術を紹介する。ベイジアンネットワークでは、用途と技術の統計的な関係をモデル化することに適用する。事前に PLSA で課題に関する用途のトピックと解決手段に関する技術のトピックを抽出し、それらの間に存在する統計的な関係性をベイジアンネットワークで分析する。

３．PLSA（確率的潜在意味解析）

３．１　PLSA の概要

PLSA (Probabilistic Latent Semantic Analysis：確率的潜在意味解析）は、文書解析のために開発された次元圧縮手法であり、教師なし学習のクラスタリング技術の一つである[11)]。人工知能

の分野では「トピックモデル」と呼ばれる技術の一つであり、テキストマイニングとはセットで適用されることが多い。

PLSA の概念図を**図 2** に示す。PLSA で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 を抽出する手法であり、この潜在クラスをトピックと呼んでいる。PLSA を文書解析で用いる場合、各文書の出現単語を記録した文書(行)×単語(列)という列数の多い高次元の共起行列データを学習し、文書 x とそこに出現する単語 y の間には潜在的な意味クラス z があることを想定して、文書と単語に共通するトピックを抽出する。これにより文書情報の全体をいくつかのトピックに集約して解釈することができる。

文書解析で用いる場合、PLSA では文書 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 (1) のように分解して考える。ここで、文書 x における単語 y の出現回数を N (x, y) とすると、式 (2) の対数尤度を最大にする P (x|z), P (y|z), P (z) を、EM アルゴリズムを用いて式 (3) の E ステップと式 (4) ～ (6) の M ステップを計算することで最尤推定する。つまり 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|z), P (y|z), P (z) の値となる。これにより「文書 x」×「単語 y」という高次元データを「文書 x」×「潜在クラス z(トピック)」という低次元データに変換することができ、文書クラスタリングの手法として用いられる

$$\mathrm{P}(x,y)=\sum_{z}\mathrm{P}(x|z)\mathrm{P}(y|z)\mathrm{P}(z) \tag{1}$$

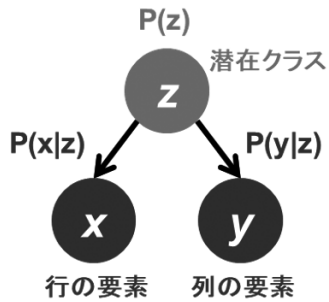
$$L=\sum_{x}\sum_{y}\mathrm{N}(x,y)\log\mathrm{P}(x,y) \tag{2}$$

$$\mathrm{P}(z|x,y)=\frac{\mathrm{P}(x|z)\mathrm{P}(y|z)\mathrm{P}(z)}{\sum_{z}\mathrm{P}(x|z)\mathrm{P}(y|z)\mathrm{P}(z)} \tag{3}$$

$$\mathrm{P}(x|z)=\frac{\sum_{y}\mathrm{N}(x,y)\mathrm{P}(z|x,y)}{\sum_{x}\sum_{y}\mathrm{N}(x,y)\mathrm{P}(z|x,y)} \tag{4}$$

$$\mathrm{P}(y|z)=\frac{\sum_{x}\mathrm{N}(x,y)\mathrm{P}(z|x,y)}{\sum_{x}\sum_{y}\mathrm{N}(x,y)\mathrm{P}(z|x,y)} \tag{5}$$

$$\mathrm{P}(z)=\frac{\sum_{x}\sum_{y}\mathrm{N}(x,y)\mathrm{P}(z|x,y)}{\sum_{x}\sum_{y}\sum_{z}\mathrm{N}(x,y)\mathrm{P}(z|x,y)} \tag{6}$$



xとyの共起確率を潜在クラスzを使って表現する

$$P(x,y) = \sum_z P(x|z)P(y|z)P(z)$$

※条件付確率P(A | B)
事象Bが起こる条件の下で事象Aの起こる確率

図２　PLSA のグラフィカルモデル

３．２　PLSA の特長

データクラスタリングという観点から、他のクラスタリング手法と比較した PLSA の特長は以下が挙げられる。

(1)　高次元データに対応できる

一般的にデータクラスタリングというと、階層クラスター分析の Ward 法や非階層の k-means 法などが有名である。こうした従来のクラスタリング手法は、データ間の「類似度（距離）」を計算し、距離の近いデータをまとめていくが、変数の数が大量にある高次元データになるほど、全体的に距離が大きく離れて妥当な結果が得られにくくなる次元の呪いと呼ばれる問題が起きてしまう。テキストデータの分析では、Bag-of-Words の形式のように単語一つ一つを列に取るなど、超高次元データとなる傾向があるため、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分かる。一方で、PLSA は高次元の情報をできるだけ保持した形で低次元に変換する次元圧縮手法であり、変数の数が多い高次元データにも対応できる。

(2)　行の要素と列の要素を同時にクラスタリングできる

上記のような従来のクラスタリング手法は、列をベースに行をクラスタリングする、あるいは行をベースに列をクラスタリングするため、どちらか一方のみがクラスタリングの対象となる。PLSA では、潜在クラスは行の要素と列の要素の 2 つの軸の変動量に基づいて抽出され、潜在クラスに対する行の要素の所属度合いと列の要素の所属度合いは式 (4)(5) によって同時に計算される。「文書」×「単語」という共起行列データをインプットとする場合、各潜在クラスのトピック

は文書と単語の両方で構成され、文書群と単語群の所属確率が計算される。つまり、一つのトピックに対して文書と単語が同時にクラスタリングされているということになる。このように、潜在クラスは行の要素と列の要素が同時に所属し、行と列の 2 つの軸の情報を持つことができるため、従来よりも情報量が多いクラスタリング結果となり解釈がしやすくなる。

(3) ソフトクラスタリングできる

上記のような従来のクラスタリング手法はハードクラスタリングと呼ばれ、ある要素が一つのグループに所属してしまうと他のグループには重複して所属が許されない。一方、PLSA はソフトクラスタリングと呼ばれ、全ての要素が全てのクラスにまたがって所属し、その各所属度合いが P(x|z) と P(y|z) で確率的に計算される。これにより、複数の意味を持つ要素がある場合でも柔軟なクラスタリングが実現できる。

3.3 トピックモデルの関連手法

トピックモデルには他に LSA (Latent Semantic Analysis)[12] や LDA (Latent Dirichlet Allocation)[13] が知られている。LSA は特異値分解によるトピックモデルであるが、テキストデータ分析でよく用いられる数量化Ⅲ類やコレスポンデンス分析も特異値分解によって軸を抽出する手法であり、LSA は数学的にはこれと同様である。なお、LSA を確率的に処理し発展させたものが PLSA となる。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) で表記しているが、数学的な考え方は同じといえる。LSA は入力する行列の成分をそのまま使用すると、大きな値をとりやすいベクトルに引っ張られて潜在クラスが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。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る。

LDA は PLSA をさらに拡張させた手法として開発されている。個々の文書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るが、LDA ではディリクレ分布という確率分布を仮定して生成させる。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、新しい文書におけるトピックの生成確率は定義されないが、LDA ではこれを推定できる。情報検索の分野では、新しいデータがどのトピックに分類されるのかということが重要となるため、PLSA よりも LDA が適用されることが主流である。

本節でテーマとしている特許文書分析において PLSA を用いる理由は、特許文書で記載されている技術の現状を把握するためである。確かに 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が、観測データのみからその現状を示す潜在クラスを抽出できる。LDA ではディレクレ分布を仮定していることで、オーバーフィッティングは回避しているが、その分

純粋な観測情報から得られた結果とは言い難いものとなっており、また抽象度が高い結果となりやすい。本節における特許文書分析では、特許文書データからその技術領域の現状における動向や、自社技術と他社技術の特徴を理解し、企業の技術戦略の検討に活用することを想定したものである。したがって特許文書データから得られる技術情報の現状をありのままに理解することが重要であると考え、本節ではトピックモデルの中でもPLSAが適した手法であると考え採用している。

３．４　PLSAの共起行列構成の工夫

行と列を同時にクラスタリングできるPLSAでは、行と列は双方が十分意味を持つ情報で構成すれば、抽出された潜在クラスの意味を2つの軸から解釈することができる。一般的なPLSAの適用では、「文書」×「単語」という構成の共起行列をインプットとするが、行に設定された「文書」はつまりは文書IDでありそれ自体に意味は持たないため、抽出された潜在クラスの意味解釈には使用しにくい情報である。また、「文書」×「単語」の共起行列は基本的に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)なデータであるため、文書間・単語間で差が出にくく、クリアで特徴的な潜在クラスが得られにくい。

こうした共起行列の構成を工夫することで解釈のしやすい潜在クラスを抽出する試みがある。一般的なPLSAにおける共起行列の構成と工夫された共起行列の構成を比較したものを**図３**に示す。例えば、「品詞Aの単語」×「品詞Bの単語」の共起行列を用いる方法が提案されており、有用な知識が抽出されたことが報告されている[14, 15]。また特許文書分析ではないが、全国の観光地の口コミから得られた「観光地」×「係り受け」の共起行列にPLSAを適用することで観光地のテーマを抽出している例もある[16]。共起行列の軸の一方を「係り受け」とすることで、文脈をイメージしやすく潜在クラスの解釈がより容易になったとされており、「単語」×「係り受け」の共起行列にPLSAを適用した事例も報告されている[17]。これにより単語と係り受けを同時にクラスタリングすることになるが、単語という話題の観点となる軸と、その観点の具体的な内容となる係り受け表現を同時にグルーピングでき、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潜在クラスとして抽出できるとされている。またこうした共起行列は単語や係り受けの出現有無に関する1か0のデータではなく、具体的な共起頻度が値として入っているクロス集計型の行列であるため、スパース性の問題の影響を受けにくく、よりまとまりのあるクリアな潜在クラスが抽出されることが期待できる。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)とても小さくなっており、計算時間も大幅に削減できる効果がある。

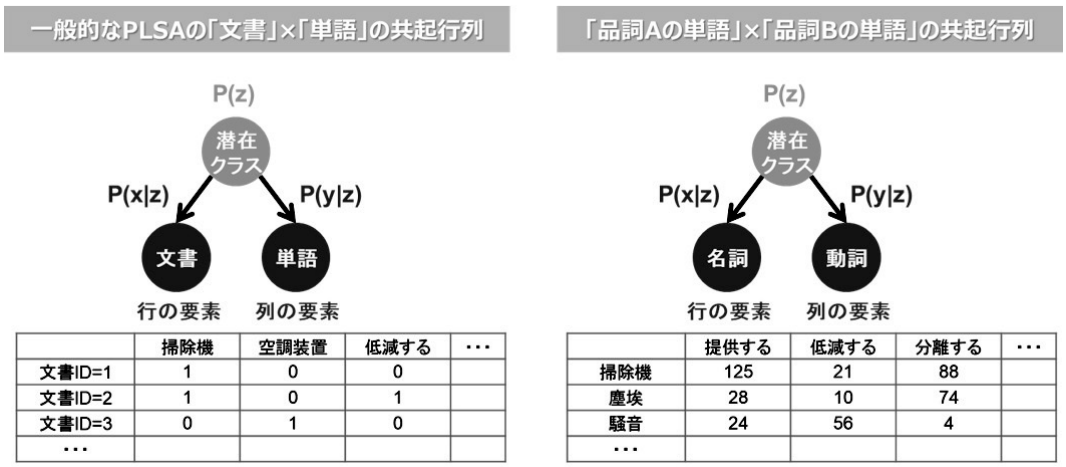


図３　PLSA の共起行列構成の工夫

４．ベイジアンネットワーク

４．１　ベイジアンネットワークの概要

ベイジアンネットワークは、複数の変数の確率的な因果関係を有向リンクのネットワーク構造で表わし、その関係の強さを条件付確率で表現した確率モデルであり、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うる確率を推論することができる[18]。ベイジアンネットワークの概要図を**図４**に示す。ベイジアンネットワークは、①確率変数、②確率変数間のリンク構造、③各リンクの条件付確率表の３つによって定義される。ベイジアンネットワークは全ての確率変数の同時確率を各変数間のリンク関係が示す条件付確率で表現するが、そのリンク構造は、観測データと変数の定義に基づいてそれを数学的に最もよく説明するモデルを学習して獲得される。これにより各変数間の確率統計的な関係性を把握することができる。また、構築されたモデルを用いることで、観測した変数群から未観測の変数の確率分布を各条件付確率表に基づいて推論することができる。なお､ベイジアンネットワークで用いる確率変数は質的変数(カテゴリカル変数)となるため、量的変数の場合は閾値を設けて事前にカテゴリに分割する必要がある。

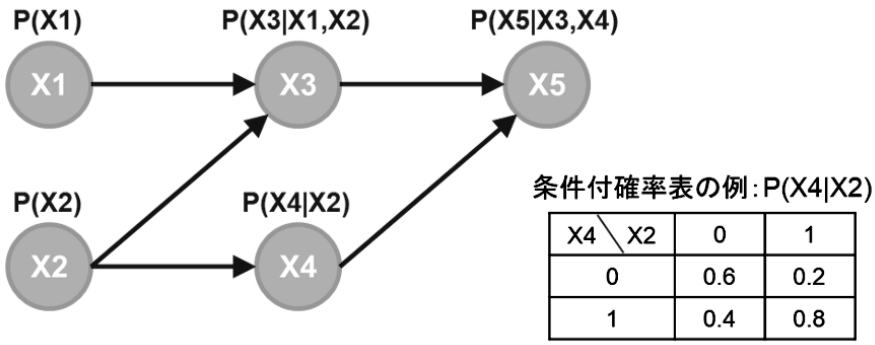


図４　ベイジアンネットワークの概要図

４．２　ベイジアンネットワークの特長

ベイジアンネットワークの特長には以下の点が挙げられる。

(1)　要因の関係構造を理解できる

本当の因果関係ではなく、あくまでも確率的な因果関係をモデル化する手法だが、どの変数がどの変数に影響しているのか、可視化された構造によってデータ全体に潜む要因関係を理解することができる。

(2)　モデルの構造を指定できる

各変数の関係構造は、観測データに基づいて数学的な基準のみで探索することもできるが、人間が構造を指定することもできる。例えばこの変数とこの変数は関係があることは分かっているといった経験則があったり、この変数群とこの変数群の関係にフォーカスして確認したいというような目的が定まっていれば、それをモデル構築の条件に採用することができ、それ以外の部分は観測データから数学的に探索するということもできる。例えば特許文書情報から用途と技術の関係を分析するという点においても、用途を実現する上での重要な要素技術を把握したいときは用途群⇒技術群という向きのリンク構造、技術を応用する用途の展開を把握したいときは技術群⇒用途群という向きのリンク構造を指定してモデルを構築すると効果的である。つまり業務における経験則や分析目的といった事前知識と数学のハイブリッドでより実務に即したモデルを構築できる。

(3)　複数の変数間の関係に基づいた様々な方向からの推論を実行できる

ベイジアンネットワークでは、目的変数と説明変数の区別なく変数の関係をモデル化し、あ

る変数を条件に与えたときの他の変数の確率を推論することができる。回帰分析や決定木分析、ニューラルネットワークなど、通常のモデリング手法では、目的変数と説明変数を設定し、一つの目的変数ごとにモデルを構築する必要があり、一つの目的変数に対する説明変数の関係しか把握することができない。また構築されたモデルを用いた推論の実行では、複数の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る。一方、ベイジアンネットワークでは目的変数と説明変数の区別がないため、それぞれの変数が互いにどのような関係性を持っているのかという構造を把握できる。また、モデルを推論で用いる場合も、その推論対象と推論条件とする変数は自由に設定でき、様々な方向から各変数の起こり得る確率を推論できる。

(4)　非線形の関係や交互作用の効果も表現できる

ベイジアンネットワークは回帰分析のように線形処理によってモデルを構築するのではなく、確率論による非線形処理のモデルのため、非線形の関係がある複雑な現象でもモデルで表現できる。また、ある条件が揃うときにだけ効果が発揮されるというものや、ある条件とある条件が組み合わさると逆の効果に転じてしまうといった交互作用がある場合でも、確率的に意味のある関係としてモデル化することができる。

４．３　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

本来ベイジアンネットワークはテキストデータを分析対象として開発された技術ではないが、これを応用することで、テキストデータの中に潜む要因関係を構造化することが可能となる。例えばテキストマイニングで抽出された単語一つ一つを確率変数に設定し、その変数間の関係をベイジアンネットワークでモデル化する取り組みが報告されている[19]。この取り組みでは病院で収集された子どもの傷害事故の診療記録データ約 4,000 件を対象に、その傷害が発生した事故状況が記されたテキスト情報にテキストマイニングを実行して事故に関わる製品や行動の単語を抽出し、それらの関係を子どもの性別・年齢の情報や生じた事故の情報と合わせてベイジアンネットワークでモデル化している。構築されたモデルを**図５**にて紹介する。このモデルによりどのような発達段階の子どもはどのような製品でどのような行動を取る可能性があり、どのような事故に至る危険性があるのかという事故発生のプロセスを確率的に推論することができる。しかし、モデルではテキストマイニングで抽出された単語一つ一つを確率変数に採用しているため、構築されたモデルはとても複雑で解釈が難しいものとなっている。これでは重要な傾向や気づきが埋もれてしまう懸念がある。そこでトピックモデルの PLSA を用いて単語をトピックに集約したものを確率変数に採用してモデルを構築することで、テキストデータ全体に存在する要因関係をよりシンプルに把握できるものと考えられる。

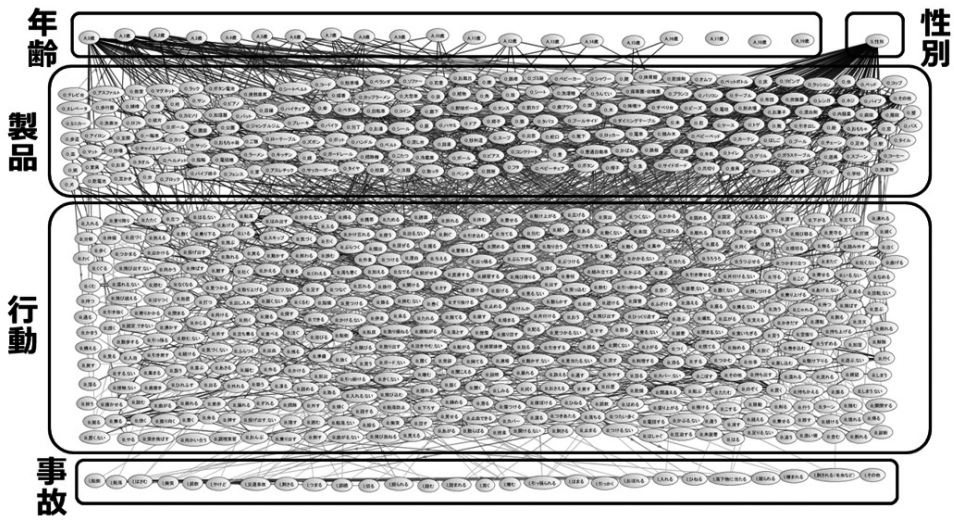


図５　子どもの傷害事故の診療記録データを用いたベイジアンネットワークモデル

（参考文献[19]より転載）

５．テキストマイニング×PLSA×ベイジアンネットワークによるテキストデータ分析手法：Nomolytics

テキストマイニングに加え、PLSAとベイジアンネットワークという2つのAI技術を適用し、テキストデータに潜む特徴や要因関係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)[20]を紹介する。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）[21]、令和4年7月10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。

５．１　Nomolyticsの手法の概要

Nomolyticsの概要を**図６**に示す。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する。次にその共起行列をインプットにPLSAを適用し、使われ方の似ている単語群をトピックにまとめ上げ、全テキストデータに対して各トピックの該当度も計算する。最後にベイジアンネットワークによってそのトピックを確率変数として扱い、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する。

こうした3つの技術を組み合わせることで、膨大なテキストデータ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、そのトピックを新たな分析軸として様々な特徴の探索が可

能となる。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る。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推論を実行することができる。

本技術はテキストデータであればあらゆる分野で適用でき、例えば旅行の口コミデータに適用して地域観光のマーケティングを検討する事例もある[17]。本節ではこれを特許文書に適用した事例について紹介する。

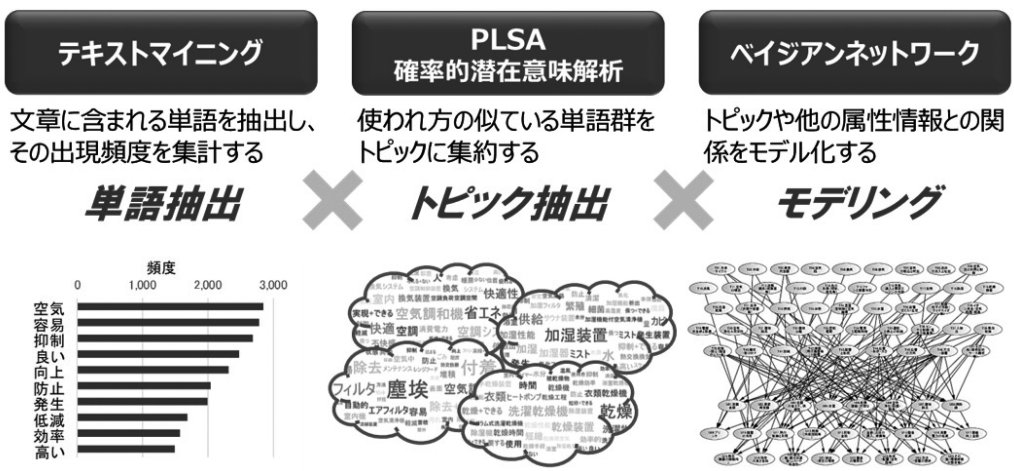


図６　PLSA とベイジアンネットワークを応用したテキストデータ分析手法：Nomolytics

５．２　各手法の連携の工夫

Nomolytics では、よりまとまりのあるクリアな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、３．４で紹介したような行列構成の工夫を施している。つまり、一般的な PLSA で適用されるような「文書」×「単語」という構成の共起行列ではなく、「品詞 A の単語」×「品詞 B 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、「単語」×「係り受け」という構成で、軸の一方を係り受けとする方法を採用している。

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するが、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る。処理の詳細は後述する実際の分析事例で解説するが、ここではその考え方を紹介する。まず、「文書」×「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各文書情報にトピックが紐づいた形で結果が得られる。そのため、この所属確率という連続値をカテゴリ化処理すれば、

各トピックは元のテキストデータに対応する確率変数としてそのまま扱うことができる。一方、Nomolytics で採用する「品詞 A の単語」×「品詞 B の単語」という構成の共起行列や、「単語」×「係り受け」という構成の共起行列で抽出されたトピックでは、各単語や各係り受けはトピックに紐づいているが (それぞれトピックに対する所属確率が計算されるが)、元のテキストデータには紐づきがされていない結果となる。つまり、元のテキストデータと抽出されたトピックの紐づきを計算する処理が必要になる。そこで、テキストデータに含まれる単語と、その単語が各トピックに対して持つ所属確率から、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する。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 1,0 のフラグに変換したトピックの確率変数を作成する。

以上のように Nomolytics では、テキストマイニングに加え、PLSA とベイジアンネットワークという 2 つの AI 技術を適用した新しいテキストデータの分析手法であるが、特にこの 2 つの AI 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る。

6．Nomolytics を応用した特許文書分析事例

ここでは Nomolytics を実際の特許文書データに適用した分析事例について紹介する。

6．1　分析の全体像

本分析事例では、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に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する。分析のプロセスの概要を**図 7** に示す。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) トレンドの分析、(4) 競合他社の分析、(5) 用途と技術の関係分析、という 5 つのステップで特許文書データを分析する。それぞれの概要は以下の通りとなる。

(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する。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する。

(2)　トピックのスコアリング

分析対象とした全特許データに対して抽出したトピックのスコア (該当度) を計算する。ここで得られた結果により、抽出したトピックを新たな分析軸とする。

(3)　トレンドの分析

各トピックのスコアを「出願年」で集計してトレンドを分析することで、用途や技術のトレンドを把握する。ここで得られた結果により、有望なニーズやシーズを探る。

(4)　競合他社の分析

各トピックのスコアを「出願人」で集計することで、各出願人の傾向やポジショニングを分析する。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、他社との提携戦略、自社技術の売却先などを検討する。

(5)　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、用途と技術の関係性を分析する。ここでの分析は、①用途⇒技術の分析（用途に対する技術の関係分析）と②技術⇒用途（技術に対する用途の関係分析）という 2 つのパターンがある。

①用途⇒技術の分析では、ある検討中の用途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ある。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、どの会社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する。

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ある。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する。

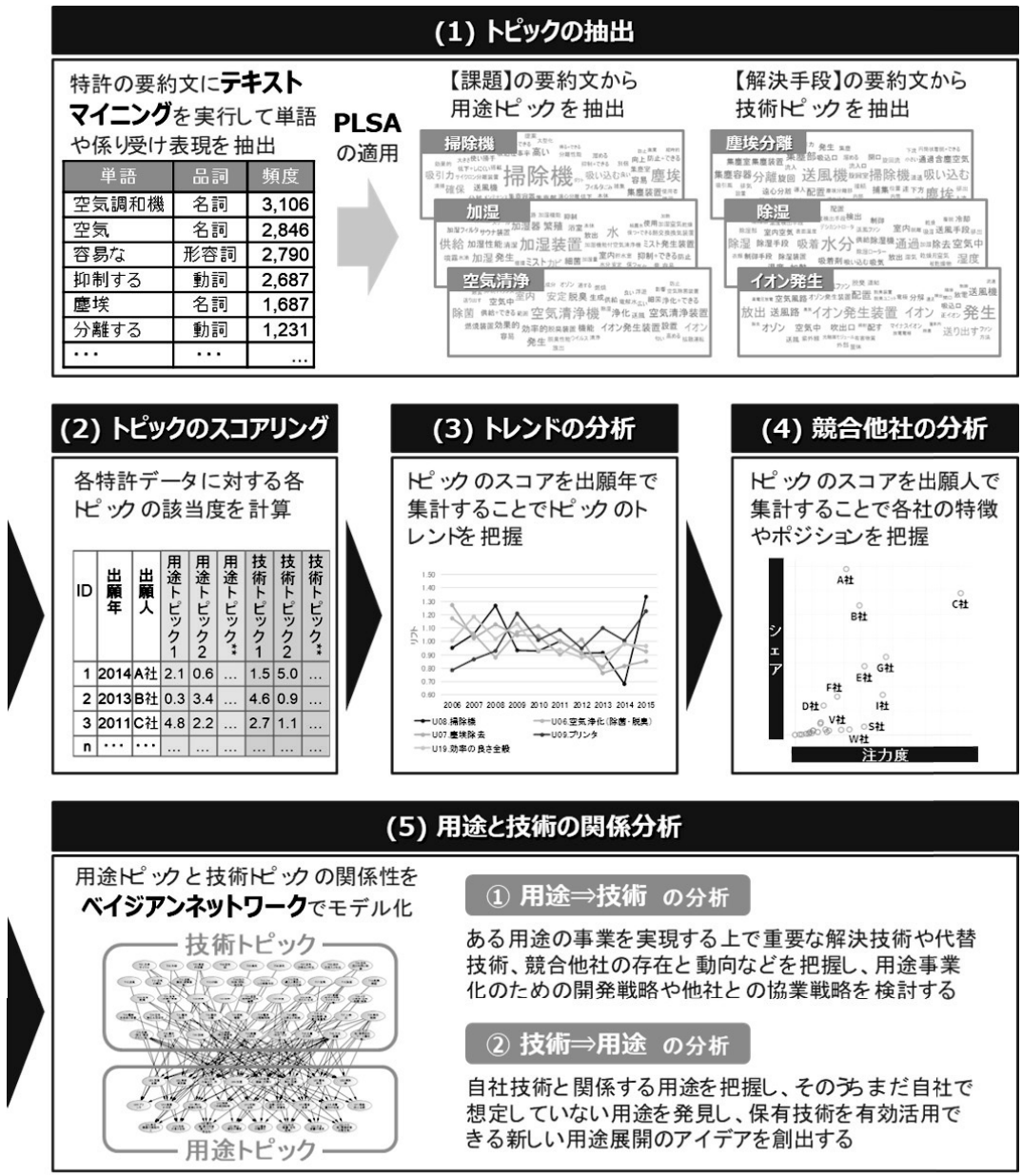


図７　Nomolytics を応用した特許文書分析のプロセス

６．２　分析で用いるデータ

本分析事例では、要約と請求項に「風」「空気」という 2 つのキーワードを含む国内の特許公報データ 30,039 件を分析対象とした。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

するため、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。「風」「空気」をキーワードに含む特許ということで、例えばエアコンや扇風機、空気清浄機、加湿器、掃除機、洗濯乾燥機など様々な生活家電が関連した特許が含まれる。

６．３　PLSA の適用による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する。国内特許の要約文では、【課題】と【解決手段】という 2 つの項目が記載されていることが多いが、【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する。その手順を以下に述べる。

(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、単語とその文法的なペアとなる係り受け表現を抽出する。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・形容動詞の単語ペアを抽出する。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）を使用している。

(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する。先述した通り、一般的な PLSA では、文書（行）×単語（列）という構成の共起行列を用いるが、本分析事例で適用する Nomolytics では、単語（行）×係り受け（列）という構成で、それぞれの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る。なお共起行列の構成に採用する単語と係り受けは頻度 10 件以上を対象とし、「課題」の文章からは単語 (3,256 語) ×係り受け (2,084 表現) の共起行列を、「解決手段」の文章からは単語 (5,187 語) ×係り受け (7,174 表現) の共起行列を作成した。なお、「提供」や「備える」など、多くの特許で頻出する単語で、かつトピックを構成する上で重要な意味を持たないと考えられる単語は、ストップワードとして除外した。

(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する。「課題」の共起行列からは用途のトピックを、「解決手段」の共起行列からは技術のトピックを抽出する。なお PLSA は予めトピック数を設定する必要があり、また与える初期値により解が異なる初期値依存性がある。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変えて PLSA を 5 回ずつ実行し、それぞれの解を情報量基準 AIC で評価して最も評価の良い解を採用する。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[22]という PLSA を拡張させた同様の分析機能を使用している。

６．４　トピック抽出の結果

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 3 つの確率が計算される。抽出された用途と技術のトピックの内容の例を**表１**に示す。なお、単語と係り受けは所属確率の高い順に並べている。**表１（左）**の用途トピック U04 では、単語は、加湿装置、水、供給、加湿、カビなどで所属確率が高く、係り受けは、加湿装置の提供、加湿器の提供、ミスト発生装置の提供、水の供給、細菌の繁殖といった表現で所属確率が高い。つまり、この結果は加湿に関するトピックであると解釈できる。**表１（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、集塵部などで所属確率が高く、係り受けは、塵埃の分離、分離する塵埃、塵埃を含む、吸い込む塵埃、含む空気、空気の分離といった表現で所属確率が高い。つまり、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る。このように「単語」×「係り受け」という構成の共起行列をインプットに PLSA でトピックを抽出することで、行と列の両方の情報軸が十分な意味を持ち、特に「単語」という話題の観点となる軸と、その観点の具体的な内容となる「係り受け」という軸でトピックが構成され、解釈のしやすい結果を得ることができている。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２**、**表３**に示す。

表１　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２　用途トピックの一覧

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（衣類など）	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３　技術トピックの一覧

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		

６．５　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) を計算する。スコアの計算では、1 件の特許の要約文章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する。なお、文 S におけるトピック T のスコアは P (S|T)/P (S) で定義する。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ある。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は高くなる。以下にこの指標の中の P (S|T) と P (S) の計算方法について説明する。

P (S|T) については、文 S を単語 X で定義される文 Sx と係り受け Y で定義される文 Sy に分解し、それぞれについて P (Sx|T) と P (Sy|T) を計算し、それらを一つに統合して P (S|T) を計算する。P (Sx|T) と P (Sy|T) はそれぞれ式 (7) と式 (8) で計算される。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 (7) の P (Sx|X) は N (X) の逆数、式 (8) の P (Sy|Y) は N (Y) の逆数として計算される。式 (7) の P (X|T) と式 (8) の P (Y|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する。そして P (S|T) は式 (9) で計算され、P (S|Sx) と P (S|Sy) は文 S において重みは同じであるためそれぞれ 0.5 とする。また P (S) は式 (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する。

$$P(Sx|T) = \sum_X P(Sx|X)P(X|T) \tag{7}$$

$$P(Sy|T) = \sum_Y P(Sy|Y)P(Y|T) \tag{8}$$

$$P(S|T) = P(S|Sx)P(Sx|T) + P(S|Sy)P(Sy|T) \tag{9}$$

$$P(S) = \sum_T P(S|T)P(T) \tag{10}$$

以上から P (S|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する。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 1,0 のフラグ情報を付与する。なお、P(S|T)/P(S) で定義したスコアは本来 1 が基準の目安となる。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あり、トピック T と文 S の間に関係があると判定できる。本分析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、その閾値を基準の 3 倍と厳しく設定した。

以上の計算処理により、**表４**に示すようなデータが作成された。30,039 件の特許データには、出願年、出願人、要約文という情報が元々あるが、そこに加え、用途トピック 25 個、技術トピック 47 個の 1,0 のフラグ情報が付加されたデータとなる。このデータセットを用いることでトピックを軸にした様々な分析を実行することができる。このデータセットを作成できればあとは従来のパテントマップと同様の分析アプローチを取ることができる。つまり PLSA でトピックを抽出しそのスコアを計算するということは、パテントマップにおいて、特許文書の記述情報を示す新たな分析軸を追加するということになっている。なお６．６以降に説明する各分析は全てこのデータセットをベースとしている。

表４　トピックのスコア（フラグ情報）を紐づけた特許データ

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	空気調和機の高	吸気口から導入	1	0		0	0	1		0
2	特願2009-XX	2009	B社	短時間で除霜を	着霜検出手段が	0	1		0	1	0		0
3	特願2011-XX	2011	C社	乾燥運転が中断	通風路を通して	0	0		1	1	0		0
4	特願2013-XX	2013	D社	ウインドシールド	車両用空調装置	0	1		0	0	1		1
…	…	…	…			…	…		…	…	…		…
30,039	特願2012-XX	2012	Z社	プリ空調時に、除	冷暖房空調ユニ	0	1		0	1	1		0

６．６　トピックのトレンド分析

表４のトピックのフラグデータを用いて、出願年の情報と、トピックのフラグ情報から、各トピックのトレンドを分析する。ここで得られた結果により、有望なニーズやシーズの探索に活用することができる。具体的には出願年 Yr とトピック T の関連度を示す指標として P(Yr|T=1)/P(Yr) を計算し、この値の経年変化を可視化する。用途トピックと技術トピックにおいて、2013 年からの直近 3 年で上昇率が高い上位 5 つのトピックのトレンドを**図８，９**に示す。

図８より、用途トピックでは、特に「U08. 掃除機」が上昇しており、それに関連してなのか「U06. 空気浄化（除菌・脱臭）」や「U07. 塵埃除去」も上昇している。一方**図９**より、技術トピックでは、「T32. 塵埃分離」や「T14. 光の利用（照射、発光等）」、「T19. 乾燥機能」、「T16. 空気導入・車両エンジンの冷却」に関する技術が上昇している。用途で掃除機のトピックが大きく上昇しており、技術では塵埃分離のトピックが上昇しているということは、直近ではサイクロン掃除機の需要と開発がホットであったと考えられる。

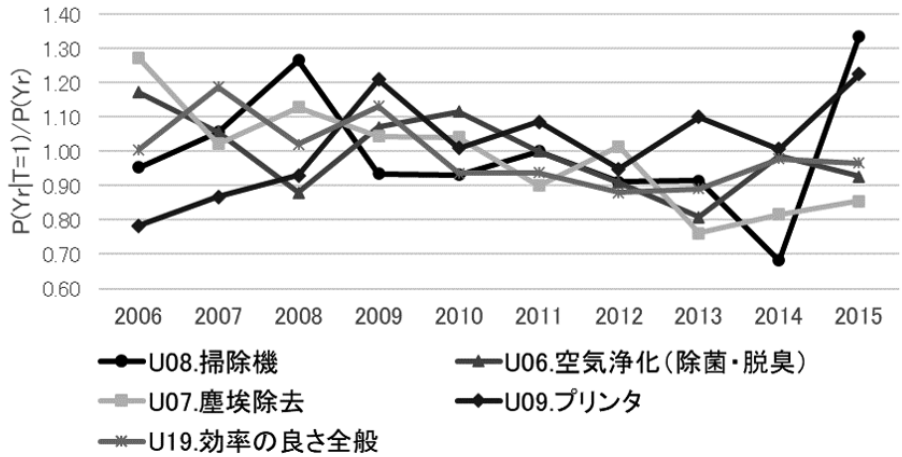


図８　2013 年から上昇率トップ 5 の用途トピックのトレンド

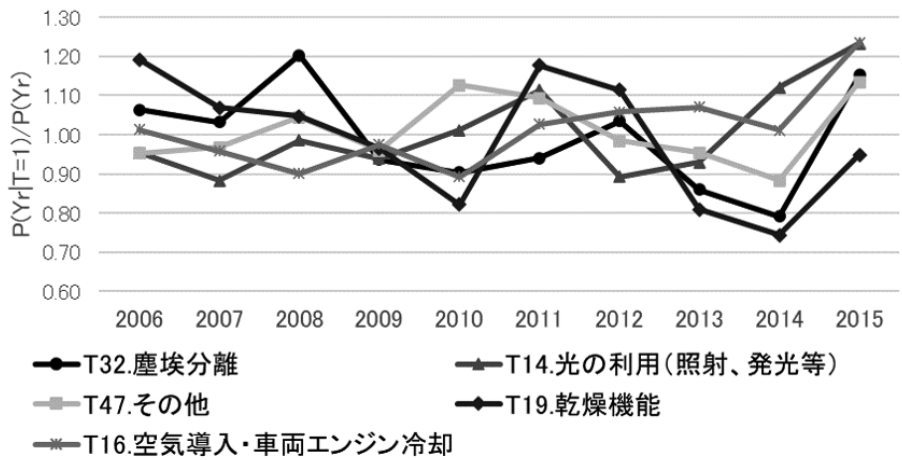


図９　2013 年から上昇率トップ 5 の技術トピックのトレンド

６．７　トピックの競合分析

表４のトピックのフラグデータを用いて、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する。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、自社技術の売却先候補などの検討に活用することができる。

本分析事例では、トピックにおける各出願人のポジショニングを可視化する。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 2 つの指標を計算し、縦軸にシェア、横軸に注力度を設定することで、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する。シェアとは P (A|T=1) 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する。注力度とは P(T=1|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 の該当割合を示す。この値が高ければそれだけそのトピック T に全社的に注力しているということであり、独自の高度な技術を保有している可能性がある。

ここでは 6．6 のトレンド分析で上昇していた技術トピック「T32. 塵埃分離」を例とした結果を**図 10** に示す。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る。**図 10** より、例えば C 社は高水準のシェアを獲得しつつ、注力度は他社と比べてとても高く、高い技術力を保有している可能性がある。今後はよりシェアを伸ばすことで高シェア高注力度の右上のポジションを確立することができる。一方 A 社と B 社もシェアは高いが、C 社には注力度においてギャップがある。そこで例えば規模は中程度だが比較的注力度が高く、高い技術力があると思われる E 社、G 社、I 社などと連携することで、右にシフトし、C 社の上のポジションを狙うことができる可能性もある。このように塵埃分離に関する技術は、1 社の注力度が高いものの、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、またトレンドもホットであったため、今後企業連携などの動きも十分考えられる領域と考察できる。

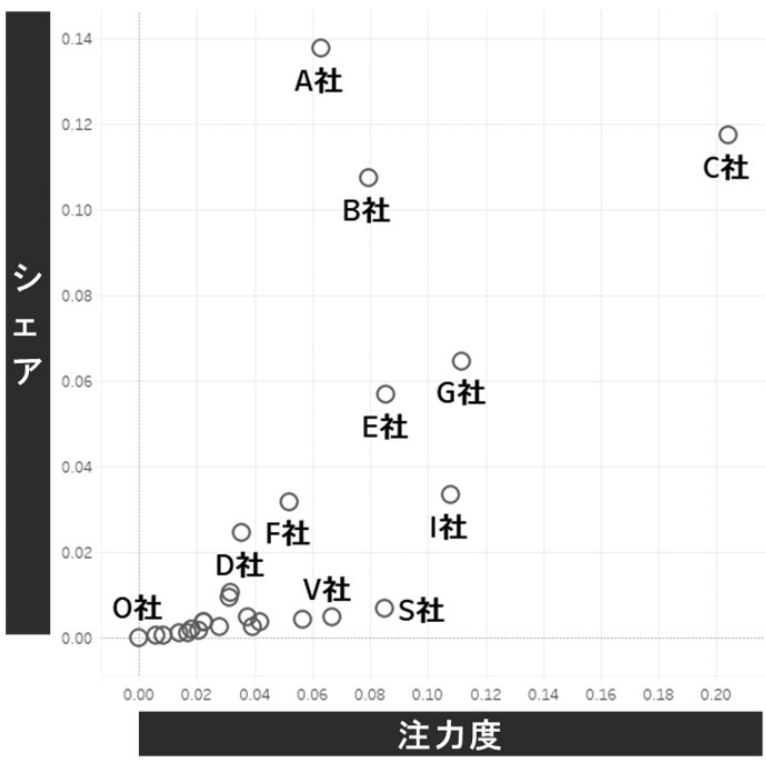


図 10　技術トピック「T. 32 塵埃分離」における出願人のポジショニングマップ

ポジショニングマップで可視化をすると各出願人の位置づけが分かりやすいが、このマップの結果はあくまでも今回の 10 年分の特許データをまとめた静的な結果であり、このマップからはその時系列の変化といった動的な傾向までは把握できない。そこで、今回注目の対象となった A 社、B 社、C 社、E 社、G 社、I 社について、この「T.32 塵埃分離」という技術トピックに該当する特許の出願件数の推移を可視化した。その結果を**図 11** に示す。

図 11 より、シェア 1 位の A 社は、近年は出願が少なく、今はあまり力を入れて開発していない可能性が考えられる。シェア 3 位の B 社は、ここ 10 年で徐々に出願が増えており、今特に力を入れている技術である可能性がある。注力度 1 位シェア 2 位の C 社は、10 年ほど前には出願件数が多く、その後一度落ち着き、直近でまた出願が急激に増えているため、再び力を入れ始めている可能性がある。E 社、G 社、I 社は、A 社、B 社、C 社と比べて全体の件数は少ないが、例えば G 社は近年出願が増えており、徐々に開発に力を入れている可能性があり、このトピックの領域において要注目と思われる。E 社は、近年では出願件数の変化が少なく、I 社は、すでに他社に買収されている会社でもあるため、2013 年以降の出願は存在していない。ここから、「T32. 塵埃分離」という技術領域では、高いシェアを誇る企業で近年競合関係にあると考えられるのは特に B 社と C 社であり、またシェアは低いものの徐々に出願を増やしている G 社の動向も今後注目すべきといえる。

このように全体でのポジショニングマップでは静的な出願人の位置づけを把握できるが、時系列での出願件数の推移も組み合わせて確認することで動的な出願人の動向も把握することができる。こうした結果から様々な技術戦略を検討するヒントが得られると期待できる。

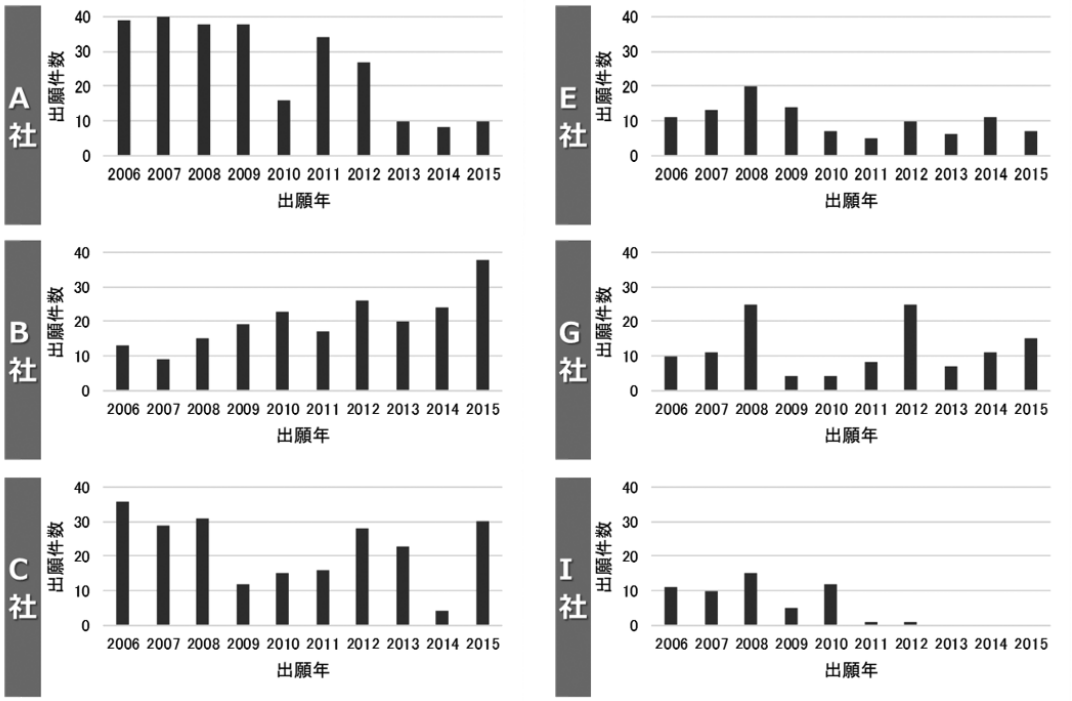


図 11　技術トピック「T.32 塵埃分離」における出願人の出現件数の推移

６.８　ベイジアンネットワークの適用による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解説する。**表４**のトピックのフラグデータを用いて、用途トピックのフラグ情報と、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する。

この関係分析は、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）という２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が逆転する。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ある。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、どの会社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の検討に活用することができる。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ある。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考えることができる。

以下にそれぞれのパターンの分析結果と考察の例を解説する。

６．８．１　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 個をリンク元に、技術トピック 47 個を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する。**図 12** に構築されたモデルの結果を示す。なおベイジアンネットワークのモデル構築には BayoLink（株式会社 NTT データ数理システム）を使用している。

図 12 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る。特に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することができるため、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することができる。ここでは６．６のトレンド分析で上昇していた用途トピック「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する。**図 12** 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 13** に示す。**図 13** のグラフでは、用途トピック U06 と関係が見られた各技術トピックの元々の確率（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（事後確率）を掲載しており、どの技術トピックも事後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分かる。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」と確認できる。

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 4 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、６．７と同様の競合分析を実施した。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た。その結果を**図 14** に示す。例えば「T26. 放電式ミスト生成」は、シェアは A 社と G 社が高いが、高シェア高注力度のポジションは空いていることが分かる。「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分かる。「T30. 塵埃吸込＆効率性」は、シェアは A 社が高いが、T26 と同じく高シェア高注力度のポジションは空いている。この結果より、例えば一強状態の技術を避けて「U06. 空気浄化（除菌・脱臭）」の用途を実現しようとするのであれば、T26 や T30 の技術が狙い目と考

えることができるかもしれない。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したり 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る。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全く新しい空気浄化家電が発売されている。つまり、その家電の技術は I 社で培われた技術であったと考察できる。

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、あるいはどの会社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することができる。

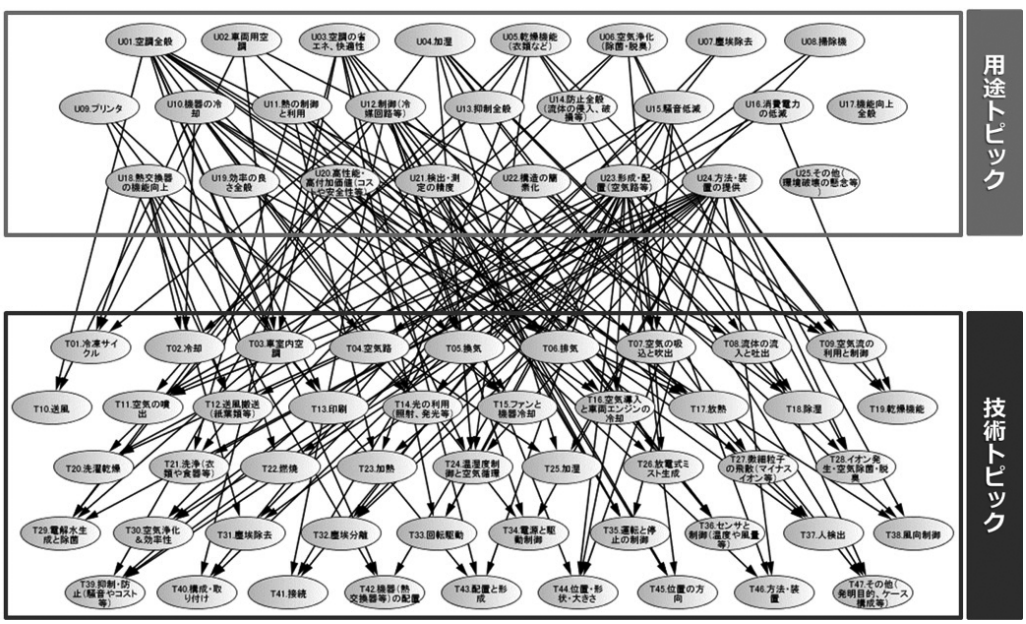


図 12　ベイジアンネットワークを適用した用途⇒技術の関係モデル

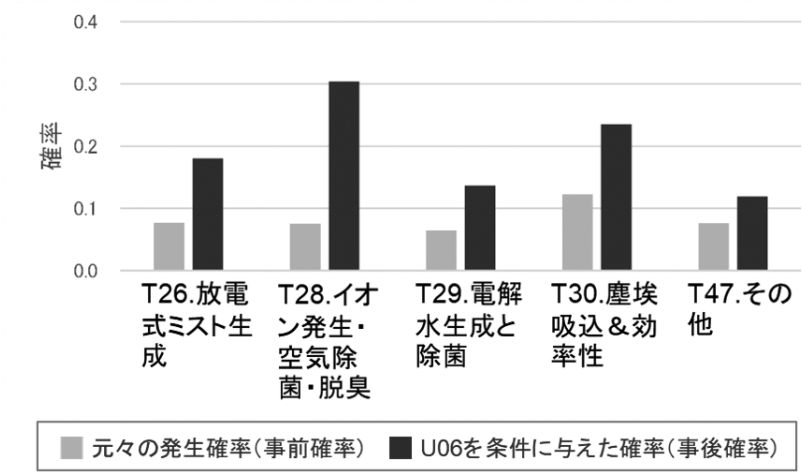


図 13　用途トピック「U06. 空気浄化」を条件に与えたときに確率が上昇する技術トピック

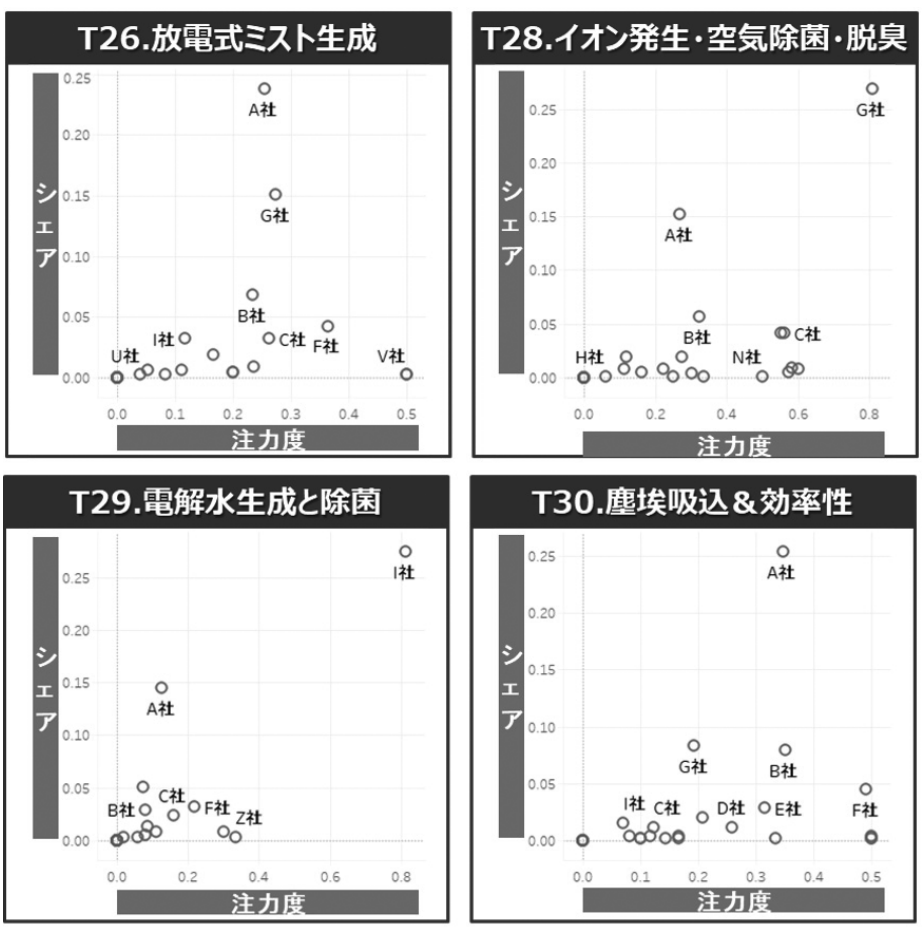


図 14　用途トピック「U06. 空気浄化」と関係する技術トピックの出願人ポジショニングマップ

６．８．２　技術⇒用途の関係分析

技術に対する用途の関係分析では、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 個をリンク元に、用途トピック 25 個を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する。**図 15** に構築されたモデルの結果を示す。

図 15 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る。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することができるため、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することができる。ここでは６．７の競合分析でも取り上げた技術トピック「T32. 塵埃分離」を対象に関係の強い用途トピックを確認した例を紹介する。**図 15** のモデルの結果より、技術トピック「T32. 塵埃分離」と関係が見られた用途トピックはただ一つ「U08. 掃除機」のみであった。ここで、技術トピック「T32. 塵埃分離」を条件に与えたときの用途トピック「U08. 掃除機」の確率を推論した結果を**図 16** に示す。**図 16** では、用途トピック U08 の元々の確率（事前確率）と、技術トピック T32 を条件に与えたときの条件付確率（事後確率）を掲載しているが、事後確率の方がとても高くなっており、「T32. 塵埃分離」の技術と「U08. 掃除機」の用途の関係が非常に強いことを確認できる。

ここではこの技術トピック「T32. 塵埃分離」と用途トピック「U08. 掃除機」の関係性に着目し、技術の新しい用途展開を探索する考え方について紹介する。まず、この両者の関係の強さを実際のデータ件数と割合でも確認すると、全体の特許 30,039 件の中で U08 の用途が該当するものは 1,559 件あり、その該当割合 P (U08) は 5.2% となる。一方で、T32 の技術が該当する特許は全体で 1,918 件あり、その中で U08 の用途が該当するものは 867 件あり、その条件付の該当割合 P (U08|T32) は 45.2% となる。つまり、T32 の技術を条件に与えることで U08 の用途の該当割合は 5.2% から 45.2% に大きく上昇するので、やはりこの技術トピック「T32. 塵埃分離」と用途トピック「U08. 掃除機」は関係がとても強いことが分かる。ところが、ある出願人 X に注目すると、X 社は T32 の技術に該当する特許が 13 件あったが、このうち U08 の用途に該当するものは 1 件も存在しなかった。つまりベイジアンネットワークで明らかになった全体における技術と用途の関係性を見れば、この X 社の保有している「T32. 塵埃分離」の技術はもっと「U08. 掃除機」の用途にも展開できる可能性があると考察することができる。

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る。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 17（左）**に示す。これはサイクロン掃除機の特許だが、特許の要約の内容を確認すると、一度集塵室内に入った塵埃が、塵埃を分離する旋回室に戻らない構成を

取ることで、集塵性能が向上し排気筒の詰まりが防止され、メンテナンスを軽減させることのできる技術として出願されている。一方、出願人X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17（右）**に示す。これは画像形成装置（プリンタ）に関する特許だが、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遠心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る。実際にX社は掃除機の製造はしていないが、X社のプリンタでトナーを分離・回収するサイクロン部の清掃時期をセンサで判断し分離効率を維持する技術は、サイクロン掃除機の集塵性能の向上にも応用できる可能性を考えることができる。

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することができる。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ないが、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となる。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ないが、こうした分析を業界の知識・経験が豊富な技術担当者が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る。

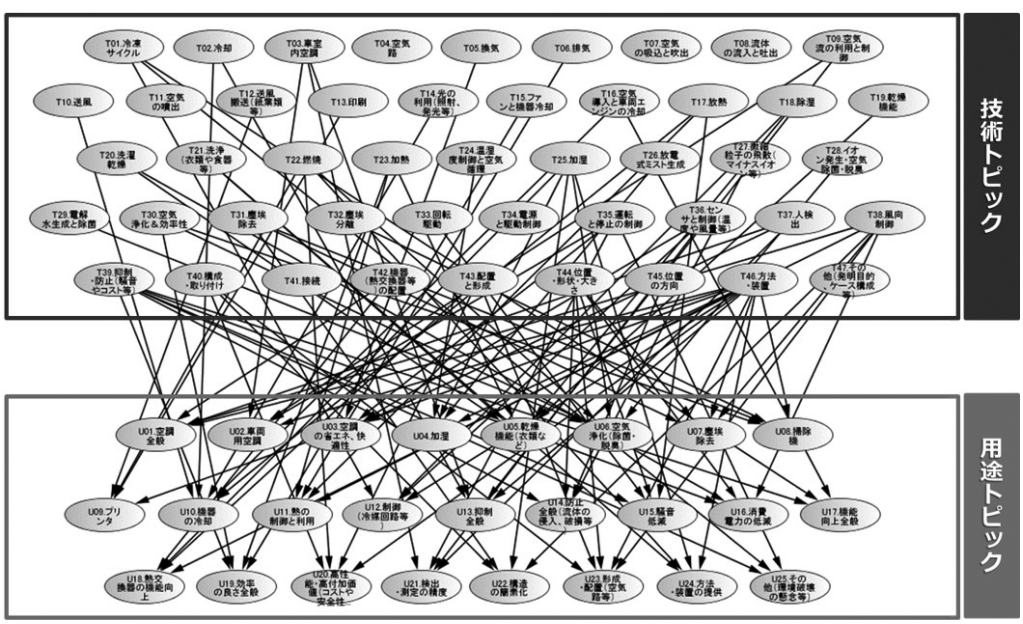


図15　ベイジアンネットワークを適用した技術⇒用途の関係モデル

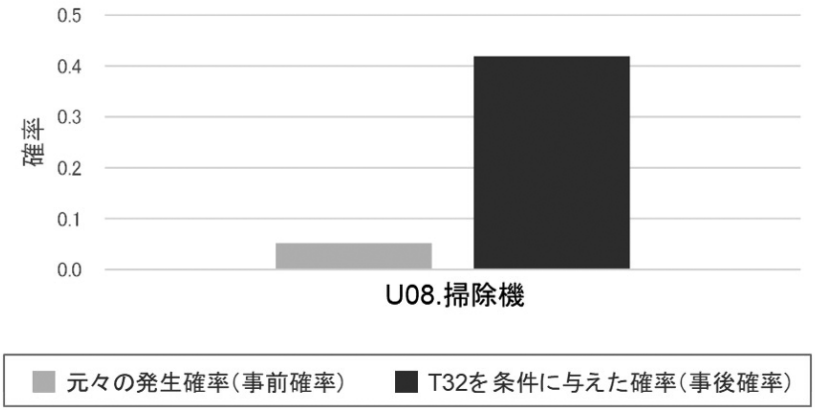


図 16　技術トピック「T32. 塵埃分離」を条件に与えたときの用途トピック「U08. 掃除機」の確率の推論結果

T32がU08に応用されている特許の代表例	出願人Xが出願しているT32に該当する特許の例
【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る

図 17　技術「T32. 塵埃分離」が用途「U08. 掃除機」に応用されている特許の 2 つの例

おわりに

本節では、特許の文書情報を用いた分析について、パテントマップと呼ばれるシンプルで分かりやすい従来の分析アプローチの有用性を認識しながら、大量の特許データを対象とする場合の分析の課題を挙げ、それを解決するアプローチとして、テキストマイニングに PLSA とベイジアンネットワークという 2 つの AI 技術を応用した分析手法 (Nomolytics) を解説し、それを実際の

特許文書データに適用した分析事例を紹介した。

Nomolytics を適用した特許文書分析のメリットには、①単語ではなく集約されたトピックを軸にした分析を実行することで、膨大な特許文書に潜む傾向を分かりやすく理解することができること、②用途と技術の統計的な関係を分析することで、各用途を実現する上で重要な要素技術を把握できたり、自社技術を有効活用できるような新規用途のアイデアを創出できることが挙げられる。

①のメリットについては、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、大量の単語をベースにした複雑な可視化アウトプットとなるため、そこから傾向を把握することが難しいという課題があった。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、カテゴリをベースに分析することもあるが、そのカテゴリの作成が属人的で作業負荷も大きいという課題もあった。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、単語ではなくそのトピックをベースにトレンドや出願人の動向を可視化することで、膨大な特許情報に潜む特徴をシンプルに分かりやすく把握することができる。特に**表４**のように各特許データに紐づけを施したトピック情報は、出願人や出願年、分類コードなどと並ぶパテントマップの新たな分析軸と見ることができる。特許文書の記述情報を集約した分析軸として、従来のパテントマップのように単純な集計をベースとしたシンプルで理解しやすい可視化を実現でき、誰もが技術戦略のための考察をすることができる。

②のメリットについては、従来の特許文書分析でも、用途と技術の関係を分析する取り組みはあったが、課題の内容と解決手段の内容に対してそれぞれ人間がカテゴリを設定し、そのカテゴリ間のクロス集計をすることで対応関係を考察するというものであり、統計的な関係までは分析できていなかった。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る。そしてその構築された関係モデルを用いることで、例えば、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や提携戦略を検討したり、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る。

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 2 つの AI 技術を組み合わせて特許文書データを分析することで、人間では読み切れない膨大な特許文書に潜む傾向や要因関係を把握でき、企業の技術戦略の検討において有益な気づきを得る新たな切り口を提供することができる。

昨今の第 3 次 AI ブームでは、特に機械学習のアルゴリズムの開発が急速に進み、技術先行型

で高度で高性能な分析アプローチが次々に提案されており、特許分析の分野でもそのトレンドは例外ではない。一方で、「データのビジネス活用」という側面では、アルゴリズムのように劇的な発展の仕方はしていないように感じる。むしろ活用面においては後退していることすら懸念される。企業では最先端のAI技術、一番いいAI技術を適用したいという思いから、より高度で難しいアプローチを採用しようとする傾向がある。しかしそれに従って分析はどんどん難解なものとなっていき、それを扱える人間は限定的となってしまう。ビジネスにおいてデータを分析する目的はデータを課題解決に活用するためであり、その最終的な課題解決の意思決定をするのは人間である。リテラシーのハードルが高すぎるとビジネスにおける活用は進みにくくなってしまう。大規模な特許データを使ってこれまで実現できなかったことを達成するような高度なアルゴリズムも重要だが、そうしたデータをビジネスに活用するという側面では、誰もがデータ分析に参加できるような単純性や操作容易性、解釈容易性も重要となってくる。そうした意味では従来のパテントマップは有用性に優れた手法といえる。古典的な集計分析ではあるものの、Excelだけでも分析と可視化が実行できる。本節で紹介したNomolyticsも、最も重要なアウトプットは**表4**に示したデータセットであり、これは元々の特許データにトピックという新たな分析軸が加わったExcelデータである。テキストマイニングとPLSAでトピックを抽出し、各データに対するトピックのスコアを計算するところまでは専門知識が求められるが、そうした専門家が**表4**のようなExcelデータさえ作成すれば、あとはそのデータを使って従来のパテントマップと同様の分析と探索を実行できる。分析は業務担当者のリテラシーに応じて自身が操作しやすく理解しやすい方法を自由に選択すればよく、例えばExcelで単純に該当件数を集計してグラフ化したり、ピボットテーブルでクロス集計したり、統計解析ソフトで分析したり、ベイジアンネットワークのようなモデリングを実行することもできる。このようにトピックという新たな分析の切り口を得たデータを用いて、誰もがデータ分析に参加しその結果を考察して技術戦略を検討できる。また抽出したトピックも、その中身の構成内容は誰もが十分解釈できるものであり、トピック抽出のベースとなっている情報は単純な単語の共起頻度だけである。

AIの技術的側面だけで見ればその発展スピードは目まぐるしく、現在人間はデータを分析する強力な武器がたくさん用意されている状況ではある。一方で、データを活用してビジネスの課題解決を達成するにはその活用的側面も評価しなければならない。今直面しているビジネスの課題の本質を明確に認識した上で、どの武器をどう使えばその課題をより効率的・効果的に解決に導くことができるのか、その武器の技術的側面だけでなくデータの活用的側面も考え、戦略的にAI技術の利用を検討することが望まれる。

参考文献

1）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.

2）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.

3）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，情報管理，Vol.52，No.1，pp.20-31，2009.

4）小池孝幸 , 石川徹也，知的財産戦略経営のための特許情報分析手法：パテントマップ作成および読解支援ツールの開発について，Japio YEAR BOOK 2008，pp.244-249，2008.

5）山中なお，知的財産戦略に資する特許情報分析事例集，特技懇，No.259，pp.82-84，2010.

6）松尾豊，人工知能は人間を超えるか ディープラーニングの先にあるもの，角川 EPUB 選書，2015.

7）Devlin, J., Chang, M.W., Lee, K., and Toutanova, K，BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding，Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies，vol.1，pp.4171-4186，2019.

8）安藤俊幸，機械学習を用いた効率的な特許調査方法―ディープラーニングの特許調査への適用に関する基礎検討―，Japio YEAR BOOK 2018，pp.238-249，2018.

9）坪田匡史，宮村祐一，神津友武，深層学習を利用した特許請求項ベースの特許技術俯瞰マップ，第 34 回人工知能学会全国大会論文集，2020.

10）難波英嗣，類似内容の特許請求項の自動対応付け，情報処理学会 第 142 回情報基礎とアクセス技術・第 120 回ドキュメントコミュニケーション合同研究発表会，2021.

11）Hofmann, T，Probabilistic latent semantic analysis，Proc. of Uncertainty in Artificial Intelligence，pp.289-296，1999.

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

13）Blei, D., Ng, A., and Jordan, M，Latent Dirichlet Allocation，Journal of Machine Learning Research，Vol.3，pp.993–1022，2003.

14) Kameya, Y., and Sato, T., Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora, Proceedings of Symposium on Large-scale Knowledge Resources, pp.65-68, 2005.
15) 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―, 第 28 回人工知能学会全国大会論文集, 2014.
16) 野守耕爾，神津友武，観光に関するユーザーレビューデータを用いた観光客の話題分析と地域観光振興への活用の検討，サービソロジー論文誌，Vol.2，No.2，pp.1-12，2019.
17) 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―観光客の声に基づいた温泉地の特徴と観光客の価値観の確率モデリング―，2016 年度人工知能学会全国大会論文集，2016.
18) 繁桝算男，植野真臣，本村陽一，ベイジアンネットワーク概説，培風館，2006.
19) 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.
20) 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.8，pp.32-337，2018.
21) 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.
22) 若杉徹，高橋勲男，医薬品調剤履歴に関する確率的構造解析に基づく適応症の推定，2014 年度人工知能学会全国大会論文集，2014.