第 5 節　特許文書分析に AI を活用する方法

(株) アナリティクスデザインラボ　野守 耕爾

はじめに

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し，自社の技術と他社の技術の特徴を俯瞰して理解することは重要である。通常，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る。特許情報を分析することは企業の技術戦略の検討において有用性が高いことは明らかである。

特許情報分析というと従来パテントマップ [1] と呼ばれるものが代表的であり，これ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，出願動向を可視化するものである。近年では特許情報の要約文や請求項，明細書などの文書を分析対象とし，テキストマイニング技術を応用することで，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よく採用されている [2]。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 [3]。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単語ベースで把握することができる。現在では複数の 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 [4]。

一方，昨今は第三次人工知能ブームと呼ばれ，コンピュータの計算性能が指数関数的に向上する中，複雑な機械学習アルゴリズムが実行可能になり，また膨大に蓄積されるビッグデータが利用可能になってきた。特に注目されている技術はディープラーニング [5] といえるが，他にも人工知能の分野で発展してきた技術には有用なものがあり，テキストマイニングによる単語の抽出と集計に留まっていた特許文書分析も，こうした人工知能技術を応用した新たな分析の検討が進められている [6,7]。

本節では，テキストマイニング技術に 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 つの人工知能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の創出が期待できる特許文書分析のアプローチとその適用事例を紹介する。

1.　従来の特許文書分析

特許文書にテキストマイニングを応用した従来の分析において，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 に示す。

1.1　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）のように単語の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。

1.2　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，単語ベースでは複雑になりすぎるため，しばしば図 1（C）のように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す

る特許件数を出願年で集計してそのトレンドを把握する。

1.3　競合他社の動向把握

他社と差別化する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候補となる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。

1.4　用途と技術の関係把握

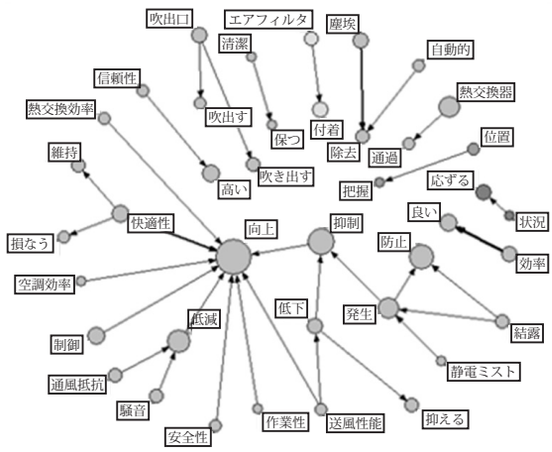
特に自社技術の新たな用途展開を探索するうえでニーズがある分析である。国内特許の要約文で記述されていることの多い「課題」と「解決手段」という 2 つの項目に着目し，それぞれの文章をテキストマイニングして図 1（C）のようにカテゴリを形成し，図 1（F）のように課題のカテゴリと解決手段のカテゴリに該当する特許件数をクロス集計することにより，用途と技術の対応関係を把握する。

1.5　従来の特許文書分析の課題

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で，①基本的に単語をベースにした分析であるため結果が複雑で考察しにくい，②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といえる。①の課題については，特にビッグデータとなるような大量の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，複雑で解釈困難な分析結果となることが多い。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。

（A）単語や係り受けの頻度集計

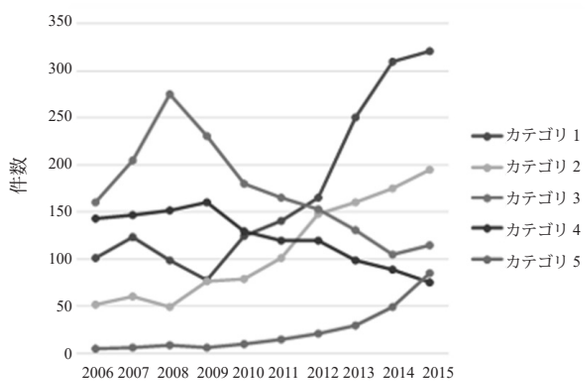
（B）単語の共起ネットワーク



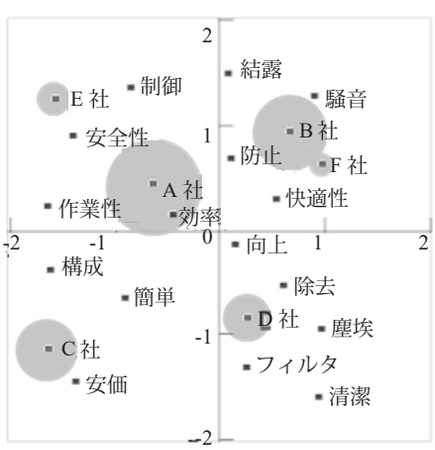
（C）カテゴリリストの作成

掃除機カテゴリのリスト	
掃除機	塵埃⇒分離
集塵	塵埃⇒吸い込む
集塵容器	塵埃⇒収容
集塵室	塵埃⇒遠心分離
吸引力	含塵空気⇒分離
サイクロン	塵埃⇒溜める

（D）各カテゴリの出願数の経年変化



（E）単語と出願人の対応マップ



（F）課題と解決手段のクロス集計

図１　従来の特許文書分析におけるアウトプット例

2. 新たな特許文書分析を展開する人工知能技術

こうした従来の特許文書分析の課題に対して人工知能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することができる PLSA（確率的潜在意味解析）という人工知能技術を，③の課題を解決する技術として，要因関係をモデリングすることができるベイジアンネットワークという人工知能技術を紹介する。

2.1 PLSA（確率的潜在意味解析）

PLSA (Probabilistic Latent Semantic Analysis) は，文書分類のために開発された次元圧縮手法である [8]。人工知能学の分野では「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。

2.1.1 PLSA の理論の概要

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，クラスタリングの手法としても適用される。開発された起源となる文書分類で用いる場合は，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，文書と単語の共通のトピックを抽出する。

文書分類で用いる場合, 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) の値である。これにより「文書」×「単語」という高次元データを「文書」×「潜在クラス（トピック）」という低次元データに変換することができ，文書分類というクラスタリングの手法としても用いられる。

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

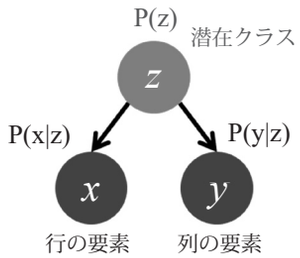
$$L=\sum\nolimits_{x}\sum\nolimits_{y}\mathrm{N}(x,y)\log\mathrm{P}(x,y) \tag{2}$$

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

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

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

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



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

2.1.2　PLSA の特長

データの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

階層型クラスター分析の Ward 法や非階層型クラスター分析の k-means 法などの従来のクラスタリング手法では, 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。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。つまり，潜在クラスに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，従来よりも情報量が多い結果となるため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ため，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できる。

2.1.3　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)[9] や LDA (Latent Dirichlet Allocation)[10] 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・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，数学的には同様の手法といえる。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である。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

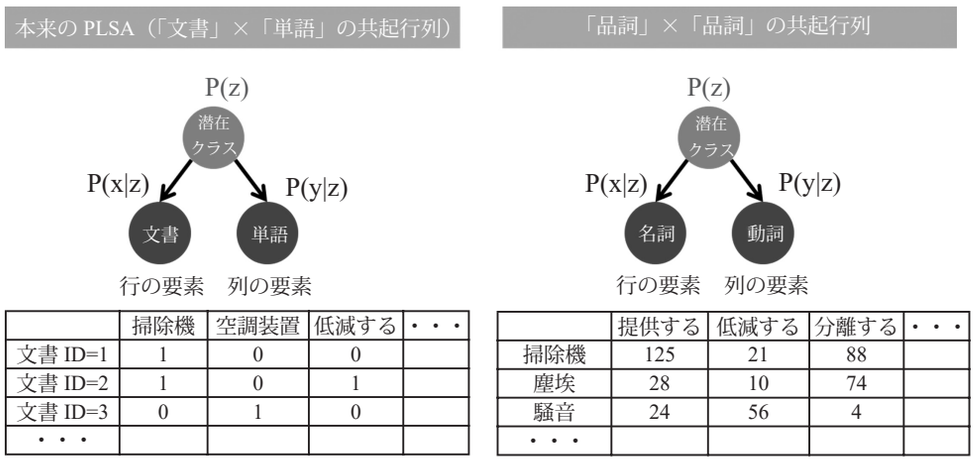
仮定して生成させる。PLSA では，観測データに過剰に適合して新規のデータの適合度が下がってしまうオーバーフィッティングが生じやすく，新しい文書におけるトピックの生成確率は定義されないが，LDA ではこれを推定できる。情報検索の分野では，新しいデータがどのトピックに分類されるのかということが重要となるため，PLSA よりも LDA が適用されることが主流である。

本節でテーマとしている特許文書分析で 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 ではディリ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，その分抽象度が高い結果となりやすく，また純粋な観測情報から得られた結果とはいえない。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俯瞰して理解し，企業の技術戦略を検討することへの活用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率直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 PLSA が適した手法であると考え取り上げている。

2.1.4　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。本来の 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 0 と 1 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特徴的でクリア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。本来の 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」×「品詞」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 [11,12]。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」×「係り受け表現」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 [13]。共起行列の軸の一方を「係り受け表現」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」×「係り受け表現」の共起行列に PLSA を適用した事例も報告されている [14]。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に基づいて，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表現の出現有無に関する 0 か 1 のデータではなく，具体的な出現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，より解釈のしやすいクリアな潜在クラスが抽出されることが考えられる。さらにその共起行列のサイズは本来の 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。



	掃除機	空調装置	低減する	・・・
文書 ID=1	1	0	0	
文書 ID=2	1	0	1	
文書 ID=3	0	1	0	
・・・				

	提供する	低減する	分離する	・・・
掃除機	125	21	88	
塵埃	28	10	74	
騒音	24	56	4	
・・・				

図 3　PLSA の共起行列構成の工夫

2.2　ベイジアンネットワーク

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 [15]。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，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。

2.2.1　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，観測データだけに基づいて数学的な基準で探索することもできるが，経験的に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事前知識や，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，それ以外の部分を観測データから数学的に探索するということができる。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技術を把握したいときは用途群⇒技術群というリンク構造，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）複数の変数を対象に様々な方向から確率シミュレーションを実行できる

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が定められており，一つの目的変数ごとにモデルを構築する必要がある。また構築されたモデルを用いたシミュレーションの実行では，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，一つのモデルで複数の推論対象を

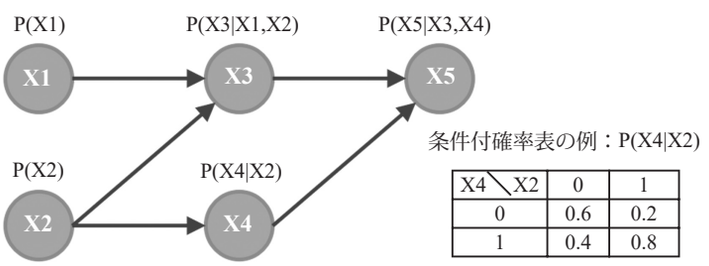
指定でき，その推論条件とする変数も区別なく自由に設定でき，様々な方向から変数の確率シミュレーションを実行できる。

（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場合でも表現できる。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。

2.2.2　テキストデータの分析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も可能である。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としてそ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[16]。この取り組みでは病院で収集された子どもの傷害データを対象に，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。これ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・傷害に至る危険があるのか確率的にシミュレーション可能にしている。しかし，モデルで確率変数に採用しているのは一つ一つの単語としているため，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おり，重要な傾向や気づきが埋もれてしまっていることも考えられる。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としてモデルを構築すれば，より全体における関係性をシンプルに把握することが期待できる。



P(X1,X2,X3,X4,X5)=P(X1)P(X2)P(X3|X1,X2)P(X4|X2)P(X5|X3,X4)

※ 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 4　ベイジアンネットワークの概要図

2.3　新たなテキスト分析技術：Nomolytics

図 5 に示すように，従来のテキストマイニング技術に加え，クラスタリング技術の PLSA とモデリング技術のベイジアンネットワークを連携させたテキスト分析技術として，Nomolytics®（Narrative Orchestration Modeling Analytics)（特許第 6085888 号）が提案されている[17]。

本技術では，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，単語間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を適用することで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因果関係をモデル化する。

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，ベイジアンネットワークによってそのテキストデータに潜む複雑な要因関係を構造化できる。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[14]。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。

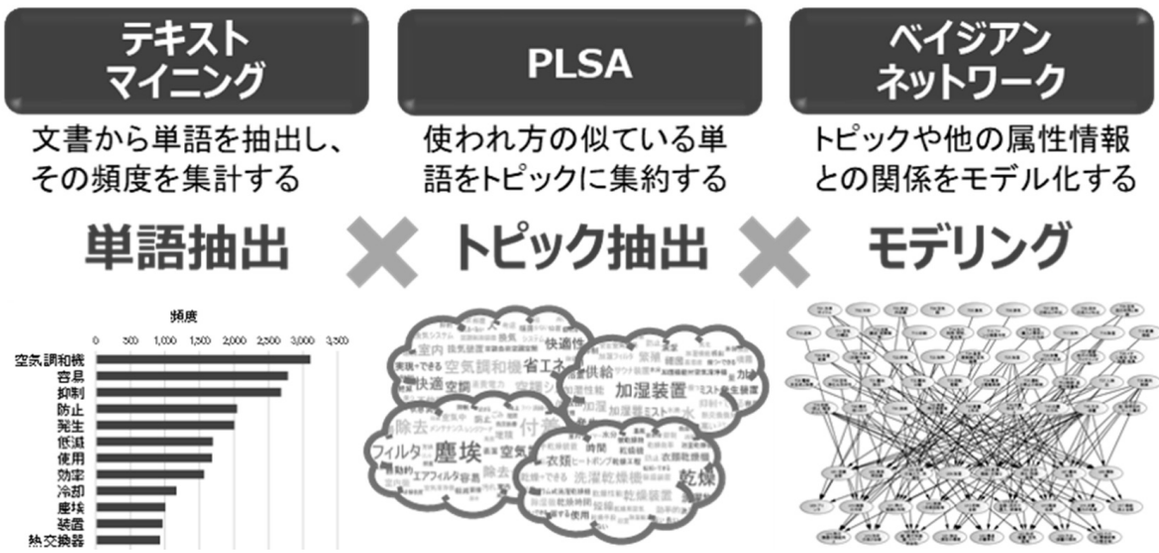


図 5　複数の人工知能技術を応用した新たなテキスト分析技術：Nomolytics

3.　Nomolytics を応用した特許文書分析事例

本節では Nomolytics を特許文書データに適用した事例について解説する。分析のプロセスの概要を図 6 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を分析する。それぞれの概要は以下の通りである。

（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。

（2）トピックのスコアリング

全ての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。

（3）トレンドの分析

トピックのスコアデータを「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し，有望なニーズやシーズを探る。

（4）競合他社の分析

トピックのスコアデータを「出願人」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，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。なお本分析は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，競合となりそうな他社はどこか，他社が牛耳る技術の代替技術は何か，どの出願人と連携すべきかなど，自社の開発戦略や他社との提携戦略を検討する。②技術⇒用途の分析では，自社技術と関係のある用途を把握し，そのうち自社で想定していない用途を見つけ，自社技術を有効活用できる新しい用途展開のアイデアを創出する。

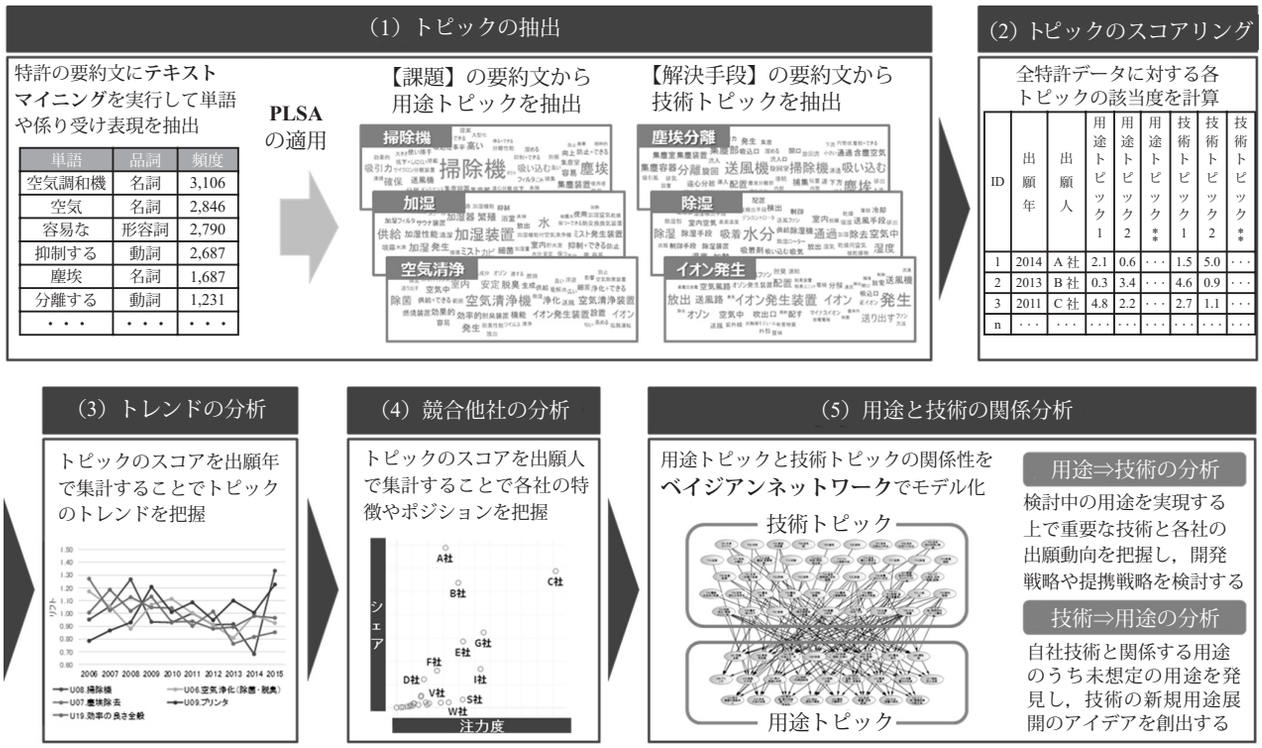


図 6　Nomolytics を応用した特許文書分析のプロセス

3.1　分析データ

本節で紹介する事例では，要約と請求項に「風」「空気」を含む 10 年分（出願日が 2006 年 1 月 1 日～ 2015 年 12 月 31 日）の国内の特許公報データ 30,039 件を分析対象とした。「風」「空気」を含む特許ということで, 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データとなっている。

3.2　用途と技術のトピックの抽出

最初に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。その方法と結果を以下に述べる。

3.2.1　トピック抽出の方法

（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，本来の 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。

（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 PLSA を初期値を変えて 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 [18] という PLSA を拡張させた同様の分析機能を使用している。

3.2.2　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。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，カビなどが，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，細菌の繁殖などが関係しているので，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，集塵部などが，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，空気の分離などが関係しているので，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る。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 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		

3.3　トピックのスコアリング

続いて全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。その方法と結果を以下に述べる。

3.3.1　トピックのスコア計算の方法

1 件の特許データには複数の文章で構成されているため，まず文章単位（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。文章 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。これはトピックを条件とすることでその文章の発生確率が何倍になるのかを示し，そのトピックをよく話題にしている文章ほど高くなる。以下，P(S|T) と P(S) の計算について説明する。

P(S|T) については，文章 S を単語で定義される文章 Sw と係り受けで定義される文章 Se に分解し，それぞれについて P(Sw|T) と P(Se|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(Sw|T) と P(Se|T) はそれぞれ式(7)と式(8)で計算される。単語 W と係り受け E が含まれる文章の数をそれぞれ n(W) と n(E) とすると, 式（7）の P(Sw|W) は n(W) の逆数, 式（8）の P(Se|E) は n(E) の逆数として計算される。式（7）の P(W|T) と式（8）の P(E|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である。P(S|T) は式（9）で計算され，P(S|Sw) と P(S|Se) は文章 S において重みは同じであるためそれぞれ 0.5 とする。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である。

$$\mathrm{P}(Sw|T) = \Sigma_W \mathrm{P}(Sw|W)\mathrm{P}(W|T) \tag{7}$$

$$\mathrm{P}(Se|T) = \Sigma_E \mathrm{P}(Se|E)\mathrm{P}(E|T) \tag{8}$$

$$\mathrm{P}(S|T) = \mathrm{P}(S|Sw)\mathrm{P}(Sw|T) + \mathrm{P}(S|Se)\mathrm{P}(Se|T) \tag{9}$$

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

以上から P(S|T)/P(S) で定義されるスコアを文章単位に計算し，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 0,1 のフラグ情報を付与する。P(S|T)/P(S) で定義したスコアは 1 が基準となるが，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，基準の 3 倍と閾値を厳しく設定した。

3.3.2　トピックのスコアリングの結果

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 0,1 の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。なお 3.4 節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	特許 2006-XXXX	2006	A 社	空気調和機の高外気温時の・・・	吸気口から導入された外気は・・・	1	0		0	0	1		0
2	特許 2009-XXXX	2009	B 社	短時間で除霜を行うことが・・・	着霜検出手段が室外熱交換機・・・	0	1		0	1	0		0
3	特許 2011-XXXX	2011	C 社	乾燥運転が中断されたとき・・・	通風路を通して回転槽内の・・・	0	0		1	1	0		0
4	特許 2013-XXXX	2013	D 社	ウインドシールドの防曇性と・・・	車両用空調装置の空調ケース・・・	0	1		0	0	1		1
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・		・・・	・・・	・・・		・・・
30,039	特許 2012-XXXX	2012	Z 社	プリ空調時に，除菌または消臭・・・	冷暖房空調ユニットは内気を・・・	0	1		0	1	1		0

3.4　トピックのトレンド分析

表 4 のトピックのスコアデータを用いて，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。具体的には出願年 Y とトピック T の関連度を示す指標として P(Y|T=1)/P(Y) を計算し，この値の経年変化を可視化する。この指標は，出願年 Y の全データにおける出願件数割合を 1 としたときに，トピック T に該当するデータにおけるその出願年の出願件数割合が何倍になっているかを示したものである。用途トピックと技術トピックにおいて，2013 年からの上昇率が高い上位 5 つのトピックのトレンドを図 7，8 に示す。

図 7 より，用途トピックでは，特に「U08. 掃除機」が上昇しており，それに関連してか「U06. 空気浄化（除菌・脱臭）」や「U07. 塵埃除去」も上昇している。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，技術でも塵埃分離のトピックが上昇しているということは，近年はサイクロン掃除機の需要と開発がホットであると考えられる。

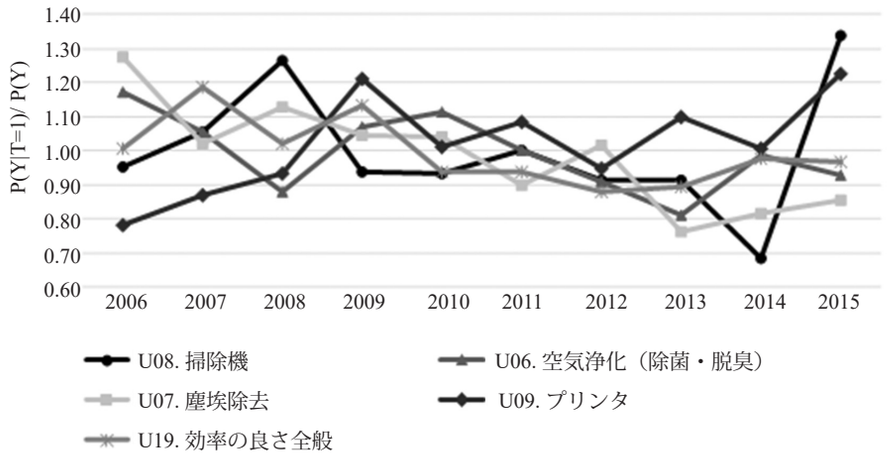


図 7　2013 年から上昇率トップ 5 の用途トピックのトレンド

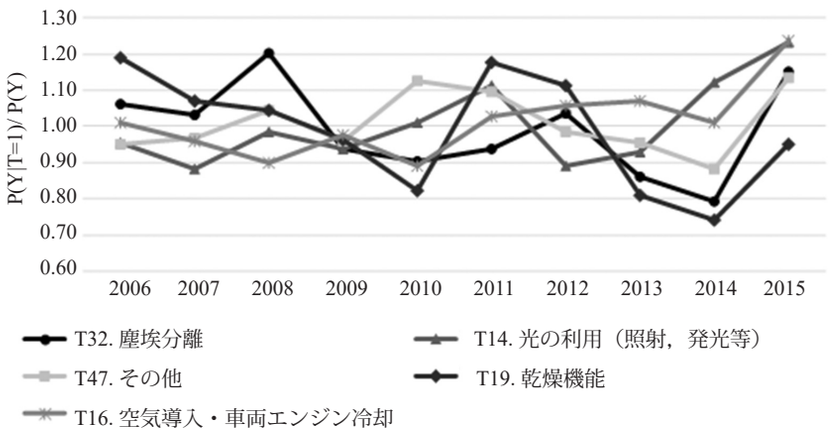


図 8　2013 年から上昇率トップ 5 の技術トピックのトレンド

3.5　トピックの競合分析

表 4 のトピックのスコアデータを用いて，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し，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。

3.5.1　出願人のポジショニングマップ

各トピックにおいて, 各出願人の位置づけを可視化する。具体的には出願人 X 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し，トピックごとに各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(X|T=1) で定義され，そのトピック T が該当する全特許の中におけるその出願人 X の出願割合を示し，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になる。注力度とは，P(T=1|X) で定義され，その出願人 X が出願した特許の中におけるそのトピック T の該当割合を示し，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，独自の高度な技術を保有している可能性がある。

本事例では 3.4 節のトレンド分析で短期的に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 9 に示す。図 9 より，C 社は高水準のシェアを獲得しつつ，注力度も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 社には注力度で劣っている。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 G 社，E 社，I 社などと連携することで，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，またトレンドも近年ホットであるため，今後企業連携などの動きも十分考えられる領域と推察できる。

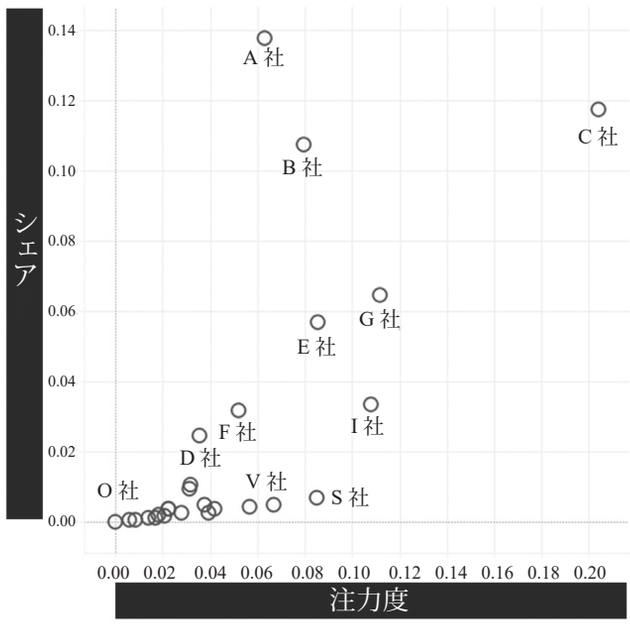


図 9　技術トピック T.32 塵埃分離における出願人のポジショニングマップ

3.5.2　出願人の出願動向

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，G 社，E 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 10 に示す。

図 10 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 年前で出願件数が多く，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。シェアが中規模であった 3 社だが，A 社，B 社，C 社と比べて件数は少ないものの，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，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，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。

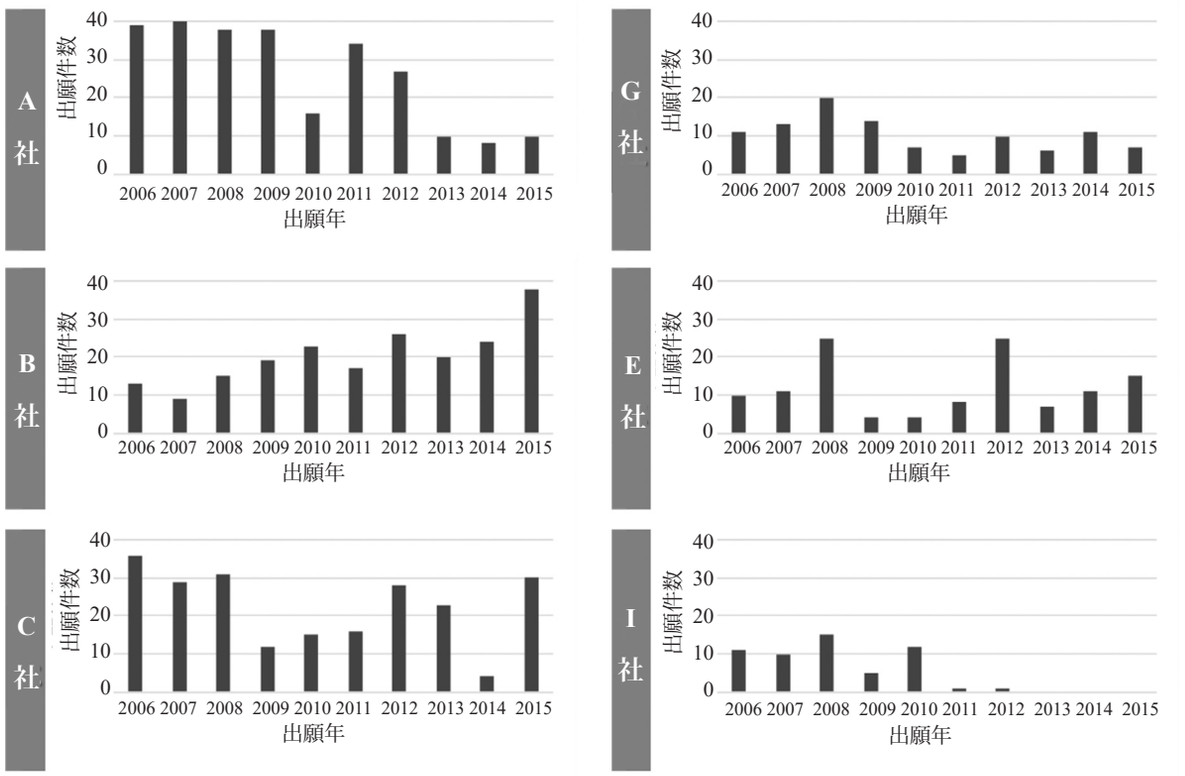


図 10　技術トピック T.32 塵埃分離における出願人の出現件数の推移

3.6　用途と技術の関係分析

表 4 のトピックのスコアデータを用いて，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から，ベイジアンネットワークを適用して用途と技術の関係性を分析する。本分析は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。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やそうして技術を保有する競合他社を把握する分析である。②技術⇒用途の分析では，自社技術と関係のある用途を把握し，そのうち自社で想定していない用途を見つけ，自社技術を有効活用できる新しい用途展開のアイデアを創出する分析である。

3.6.1　用途⇒技術の関係分析

用途に対する技術の関係分析では，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性を可視化する。図 11 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 11 のモデルを用いることで，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，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することができる。本事例では 3.4 節のトレンド分析で短期的に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。図 11 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の技術トピックの確率を推論した結果を図 12 に示す。図 12 では，用途トピック U06. 空気浄化を条件に与えた条件付確率が元の確率（何も条件を与えていない確率）よりも上昇した技術トピックのみを掲載している。図 12 より，U06. 空気浄化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」であった。

用途トピック U06. 空気浄化と関係のあった 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，3.5 節と同様の競合分析を実施した。用途トピック U06. 空気浄化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結果を図 13 に示す。「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。またＦ社はややシェアが低いが，高い注力度がある。「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，T29 は I 社が牛耳っている。「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おり，こちらも F 社がややシェアは低いが高い注力度がある。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られるが，そのなかでも注力度の高い F 社は注目となる。逆に一強状態にある T28 や T29 の技術において，その一強企業と提携あるいは 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る。なお，先述した通り「T29. 電解水生成と除菌」を牛耳る I 社はすでに買収されているが，実際にその買収した会社から電解水（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る。

このように自社で検討中の用途に関係する重要な解決技術を分析することで，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，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あるか，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，開発戦略や提携戦略を検討することができる。

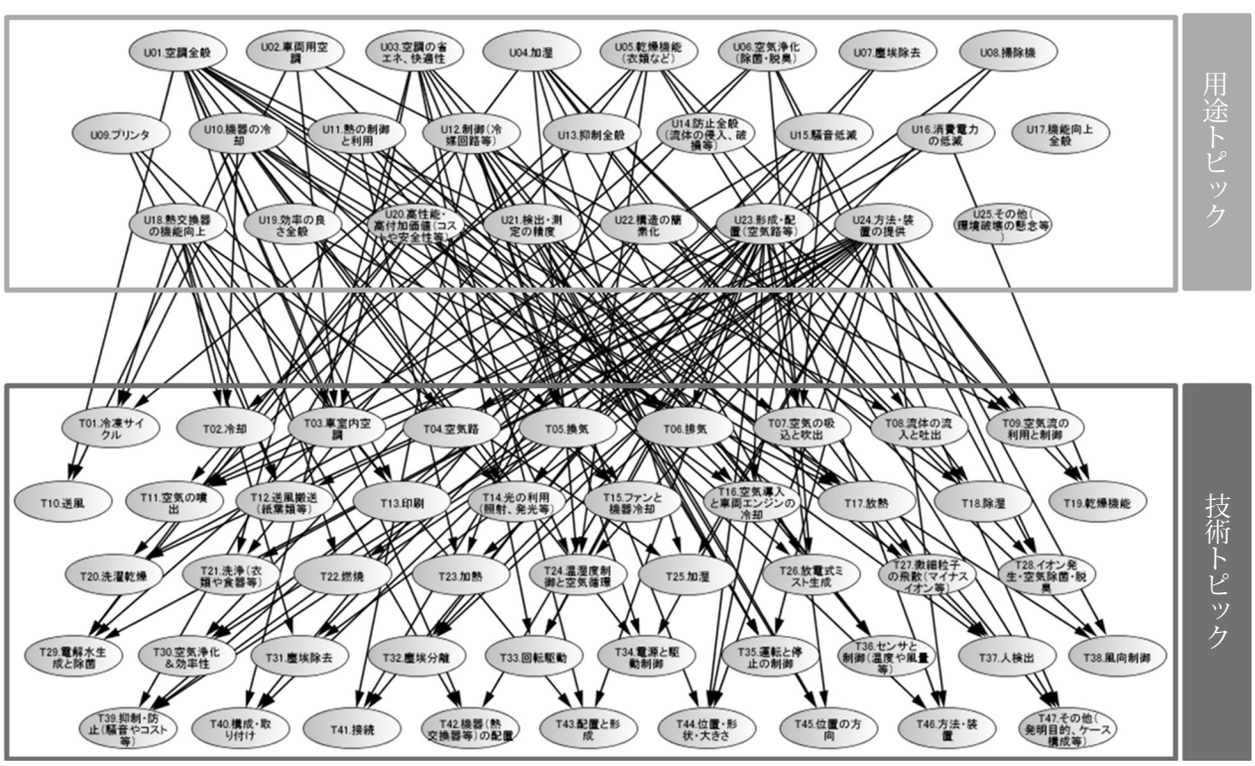


図 11　用途に対する技術の関係モデル

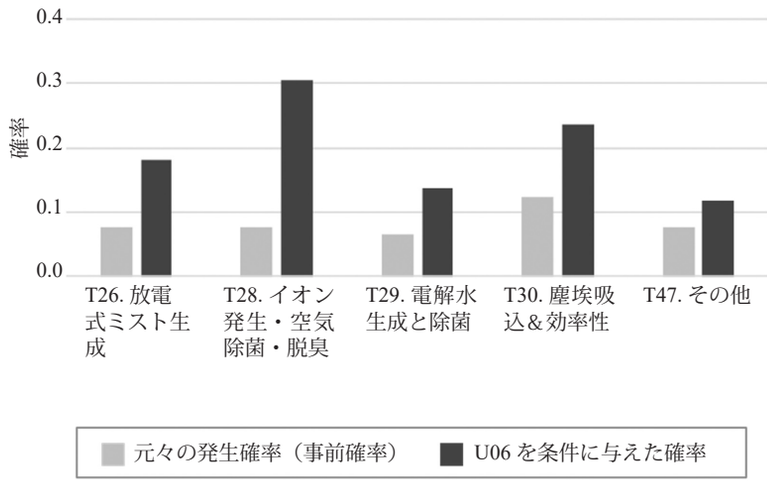


図 12　用途 U06. 空気浄化を条件に与えたときに確率が上昇する技術トピック

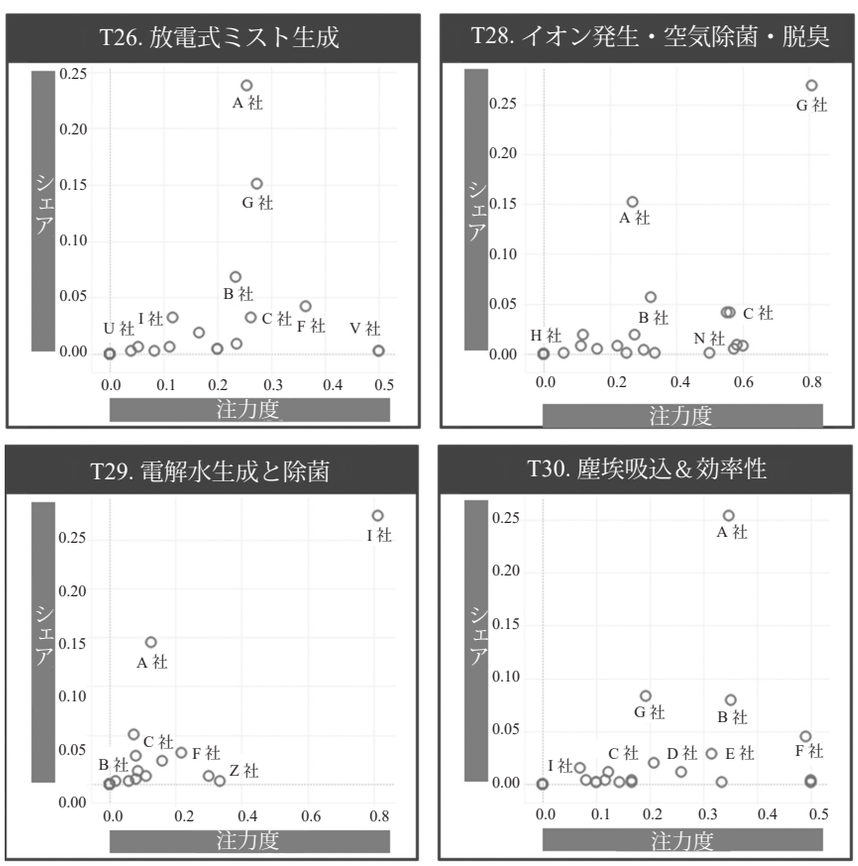


図 13　用途 U06. 空気浄化と関係のある技術トピックの出願人ポジショニングマップ

3.6.2　技術⇒用途の関係分析

技術に対する用途の関係分析では，3.6.1 節の用途⇒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性を可視化する。図 14 に構築されたモデルの結果を示す。

図 14 のモデルを用いることで，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。本事例では技術トピック「T18. 除湿」を対象に関係の強い用途トピックを確認した。図 14 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の用途トピックの確率を推論した結果を図 15 に示す。図 15 では，技術トピック T18. 除湿を条件に与えた条件付確率が元の確率（何も条件を与えていない確率）よりも上昇した用途トピックのみを

掲載している。図 15 より，T.18 除湿の技術と関係の強い用途トピックは，「U05. 乾燥機能（衣類など）」，「U12. 制御（冷媒回路等）」であったが，ここでは技術 T18. 除湿と用途 U05. 乾燥機能の関係を例に取り上げ，技術の新しい用途展開を探索する考え方について説明する。

この両者の関係が強いということは，U05. 乾燥機能の用途に該当する特許は，全体よりも T18. 除湿の技術に該当する特許の中の方が多くの割合で存在するということだが，ある出願人 X に注目すると，X 社は T18. 除湿の技術に該当する特許のうち U05. 乾燥機能の用途に該当するものはほとんど存在しなかった。つまりベイジアンネットワークによ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 T18. 除湿に関する技術はもっと U05. 乾燥機能の用途に展開できる可能性があると考えることができる。

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進めることができる。まず T18. 除湿の技術が U05. 乾燥機能の用途を想定して出願されている特許の代表例としてドラム式洗濯乾燥機の特許がある。特許文書の内容を確認すると，例えば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ための除湿技術が求められている。一方出願人 X が出願している T18. 除湿の技術に該当する特許には，インクジェットプリンタに関する特許があった。インクジェットプリンタでは，吹き付けるインク液にムラが出ないようにすること，またそのインク液を吸収した紙が湿度のムラによって波打たないように乾燥処理することが求められている。X 社はプリンタの中で，紙に残った余分なインク液を加熱して蒸発させ，その蒸発による湿気を吸引ファンで取り除くことで，インク液が不均一にならないように乾燥処理をする技術を特許として出願していた。X 社は洗濯乾燥機の製造はしていないが，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乾燥処理技術は，例えば洗濯乾燥機の中で洗濯物をムラ無く効率的に乾燥させることに応用できる可能性も考えられる。これはあくまで分析結果から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ことが期待できる。

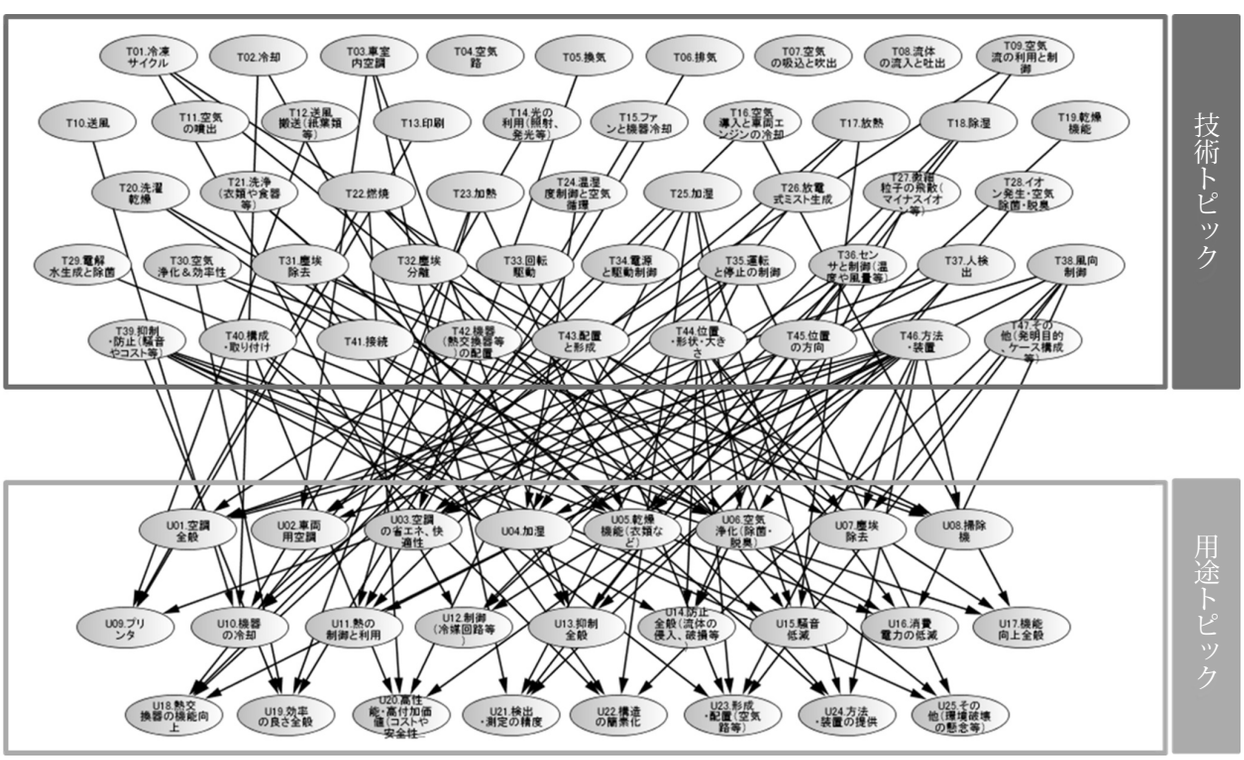


図 14　技術に対する用途の関係モデル

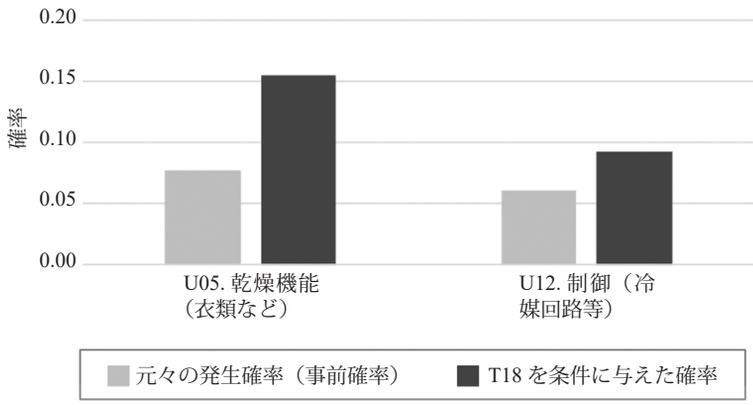


図 15　技術 T18. 除湿を条件に与えたときに確率が上昇する用途トピック

おわりに

本節では，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 つの人工知能技術を応用した新たなテキスト分析技術（Nomolytics）と，それを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ベースにした分析を実行することで，膨大な特許文書に潜む特徴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，用途を実現する上で重要な要素技術を把握できたり，技術の新規用途のアイデアを発想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分析では，単語をベースにした複雑な結果を解釈しなければいけないため特徴が把握しづらく，その単語を人間がグルーピングしてカテゴリを作成することでカテゴリベースに分析することもあるが，そのカテゴリ作成が属人的で作業負荷も大きいという課題があった。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各出願人の特徴を分析することで，膨大な特許情報に潜む特徴をシンプルに分かりやすく理解することができる。

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ことはあったが，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その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。またその分析結果を用いて，例えば，検討中の用途に対して関係の強い技術やそうした技術における出願人の動向を把握することで，自社の技術戦略や提携戦略を検討したり，あるいは自社技術と関係の強い用途を確認し，そこでまだ想定していない用途とそれに関連する特許文書を探索することで，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，人工知能技術を応用して人間では読み切れない膨大な特許文書を分析し，そこに潜む特徴や要因関係を把握することで，企業の技術戦略の検討において新たな知見の創出が期待できる。

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）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）
3）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）
4）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）
5）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）
6）安藤俊幸，機械学習を用いた効率的な特許調査方法，Japio YEAR BOOK 2016，pp.150-161（2016）

7）岩本圭介，特許文献から技術動向を把握するためのマイニング手法，Japio YEAR BOOK 2016，pp.198-203（2016）
8）Hofmann, T，Probabilistic latent semantic analysis，Proc. of Uncertainty in Artificial Intelligence，pp.289-296（1999）
9）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）
10）Blei, D., Ng, A., and Jordan, M, Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp.993-1022（2003）
11）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）
12）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）
13）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）
14）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）
15）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）
16）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）
17）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）
18）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）