第 10 節　トピックモデルを応用したテキストデータの理解とインサイトの獲得

（株）アナリティクスデザインラボ　野守 耕爾

はじめに

2012 年 3 月に当時のアメリカのオバマ政権から“Big Data Research and Development Initiative”[1] というビッグデータの研究開発プロジェクトが発表された。またアメリカのビジネス雑誌である『Harvard Business Review』の 2012 年 10 月号（日本語版では 2013 年 2 月号）では，“Data Scientist：The Sexiest Job of the 21st Century”というタイトルの論文[2] が，データ経営の第一人者であるトーマス・ダベンポート氏から寄稿され，データサイエンティストは 21 世紀で最も魅力的な職業であると表現された。こうして 2012 年頃より，ビッグデータ，データサイエンス，データアナリティクス，AI，IoT，DX，データドリブン経営といったキーワードとともに，データを分析しビジネスに活用する取り組みが企業で活発化している。そうしたなか，文章で記されたテキストのデータも活用が進んでいる。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，そこには今まで気づかなかった有用な知見が隠れている可能性が大いにある。ビジネスの課題解決において，こうしたテキストデータを分析することは，重要なインサイトの獲得が期待できる取り組みである。

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言われるが[3]，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テキストマイニングの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実施されるようになった。テキストマイニングを実行することで，テキストデータの文書内容を人間が読み込まなくても，そこに含まれる単語を抽出し，単語単位による集計と可視化に基づいて全体の記述傾向の現状把握ができる。しかし，分析対象のデータがビッグデータになってくると，テキストマイニングで抽出される単語は膨大となり，そうした大量の単語を単位とした可視化は複雑で解釈が困難となる。これではビジネスにおける課題解決の取り組みを妨げてしまう。そこで自然言語処理分野における AI 技術の一つで，使われ方の似ている単語群を集約できるトピックモデルという技術を応用することが有効である。大量の単語をベースに可視化をするのではなく，いくつかの集約されたトピックをベースに可視化をすることで，テキストデータの特徴をシンプルに理解することができるため，課題解決のための示唆が得やすくなるアプローチとなる。

本節では，テキストマイニングにトピックモデルを応用することで，より容易にテキストデータの理解を進め，より効率的にインサイトを獲得する分析アプローチとその適用事例を紹介する。なお，本節で使用する「テキストデータ」とは，「文書情報」と「属性情報」を含むデータとする。

1.　テキストマイニング

ここではテキストデータの分析と活用の意義について確認した上で，従来の分析手法となるテキストマイニングの概要とその課題について述べる。

1.1　テキストデータの分析と活用の意義

ビジネスにおけるテキストデータの分析と活用の例を挙げると，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することができるため，マーケティング戦略に活かすことができる。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，消費者の評価を競合他社の商品と比較しながら理解でき，現状商品の改善や新たな商品企画に活かすことができる。また，コールセンターに寄せられた問い合わせ内容が記録されたテキストデータを分析すれば，顧客の要望や不満の内容とその動向を理解でき，自社の商品やサービスの改善に活かすことができる。他にも，公開されている特許

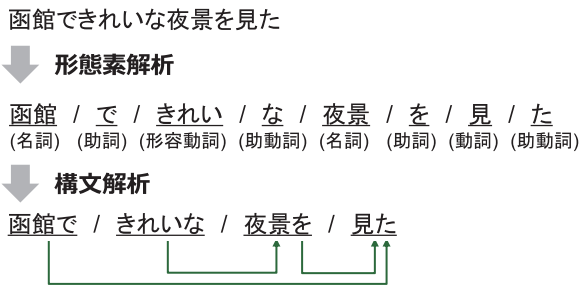
情報のテキストデータを分析すれば，競合他社の技術動向を把握し，自社の研究開発の計画や M&A などの経営戦略に活かすことができる。

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る。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るが，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」は不足している。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ととも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る。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ができる。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課題解決において有用で，価値の高い取り組みといえる。

1.2　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するとき，その文書情報は定性的なデータであるため，そのままでは他の定量的なデータのように集計したり統計解析を実行することはできない。テキストマイニングはそんな文書情報のデータを定量的に統計分析可能にするデータマイニング手法であり，大量のテキストデータからその文書に含まれる単語を抽出し，単語単位による集計や統計解析に基づいて文書に記述されている傾向を可視化する。

テキストマイニングにおける自然言語処理には「形態素解析」と「構文解析」という 2 つの基本技術がある [3]。形態素解析とは，文章を単語に分割し，その単語の品詞を割り当てる技術である。正確には単語ではなく，形態素と呼ばれる意味を持つ最小単位の文字列として分割する手法だが，実用面では，形態素ではなく単語の単位で分割されることが通例となっている。単語と形態素の違いの例として，たとえば「観光地」という単語は「観光」「地」という 2 つの形態素で，「宿泊施設」という単語は「宿泊」「施設」という 2 つの形態素で構成される。このように，日本語の意味情報を抽出するには最小単位の形態素よりも単語の方が適しているとされる。一方，構文解析とは，文法規則に基づいて文章の構造を解析し，単語間の関係性を識別する技術である。たとえば，主語と述語や，修飾語と被修飾語といった係り受け関係を抽出する。形態素解析と構文解析の例を図 1 に示す。これらの処理により，文書に含まれる単語とその品詞，また単語間の係り受け関係を抽出することができる。なお，実際に抽出される単語や係り受けは，登録されている辞書によってそれぞれ原形に変換された形で抽出されることが一般的である。形態素解析と構文解析には公開されたフリープログラムがあり，形態素解析では JUMAN，ChaSen，MeCab などが，構文解析では KNP や CaboCha などが広く知られている [4]。



形態素解析で抽出される単語と品詞

単語	単語の原形	品詞
函館	函館	名詞
で	で	助詞
きれい	きれい	形容動詞
な	だ	助動詞
夜景	夜景	名詞
を	を	助詞
見	見る	動詞
た	た	助動詞

構文解析で抽出される係り受け

函館⇒夜景
きれい⇒夜景
夜景⇒見る

図 1　形態素解析と構文解析の例

テキストデータという定性的な文書データを定量的に統計分析可能にするには，そのデータを数値の形式で表現する必要があるが，その代表的な方法が文書や単語をベクトルで表現するというものである。形態素解析はその最も基本的な技術であり，形態素解析で文書に含まれる単語を抽出し，各文書にどの単語が何回出現したのかをカウントすることで，その出現頻度という数値によって文書一つひとつをベクトルとして表現できる。このように単語の出現頻度でもって文書をベクトル化する手法を Bag-of-Words（BoW）と呼ぶ。Bag-of-Words のベクトル表現は「文書×単語」という行列形式のデータで考えると分かりやすい。一つの行に一件の文書を，一つの列に一つの単語を取ったとき，各セルの値はその文書内におけるその単語の出現回数を示す。Bag-of-Words によるデータ形式の例を図 2 に示す。これは最もシンプルな文書のベクトル表現だが，この処理だけでもテキストデータのさまざまな定量分析が可能であり，とても強力な処理手段である。テキストマイニングのツールで実行される可視化や統計解析は基本的にこのデータ形式をベースにしている。

文書ID	ホテルの口コミコメント	部屋	広い	快適	空調	良い	効く	綺麗	清掃	風呂	駅	近い	便利	スタッフ	対応	丁寧	朝食	美味しい	レストラン	…
1	部屋が広く快適で、部屋の空調も良く効きました。	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
2	部屋は綺麗に清掃されていて、お風呂も快適でした。	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	
3	駅の近くで便利で良く、部屋も広くて良かったです。	1	1	0	0	2	0	0	0	0	1	1	1	0	0	0	0	0	0	
4	スタッフの対応が良く、部屋も丁寧に清掃されていました。	1	0	0	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0	
5	朝食が美味しく、レストランも広くて綺麗で良かったです。	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	
…	……																			

図 2　Bag-of-Words の例

テキストマイニングは現在多くの実行ツールが存在し，自然言語処理の専門知識がなくても分かりやすいインタフェースで直感的な操作が可能で，テキストデータを活用したいビジネスの現場では人気のある分析ツールである。たとえば，無償のツールでは，立命館大学の樋口耕一教授が作成した KH Coder やユーザーローカル社の AI テキストマイニングなどが，有償のツールでは，NTT データ数理システム社の Text Mining Studio やプラスアルファ・コンサルティング社の見える化エンジンなどが知られている。

テキストマイニングのツールでよくある分析のアウトプット例を図 3 に示す。もっとも単純な分析は単語の頻度集計である（図 3 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書情報では，どのような記述が多いのかという全体像を把握する。また，1 件の文章の中で同時に出現しやすい単語同士をネットワークでつなぐ共起ネットワークという分析がある（図 3 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ため単語間のつながりが分からないが，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る。こうしたネットワーク結果から，このテキストデータの文書情報ではどのような話題が形成されているのかを考察できる。さらに少し高度な分析として，コレスポンデンス分析（あるいは数量化Ⅲ類）というものもある（図 3 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年や出願人といった属性情報）と文書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ある。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る。

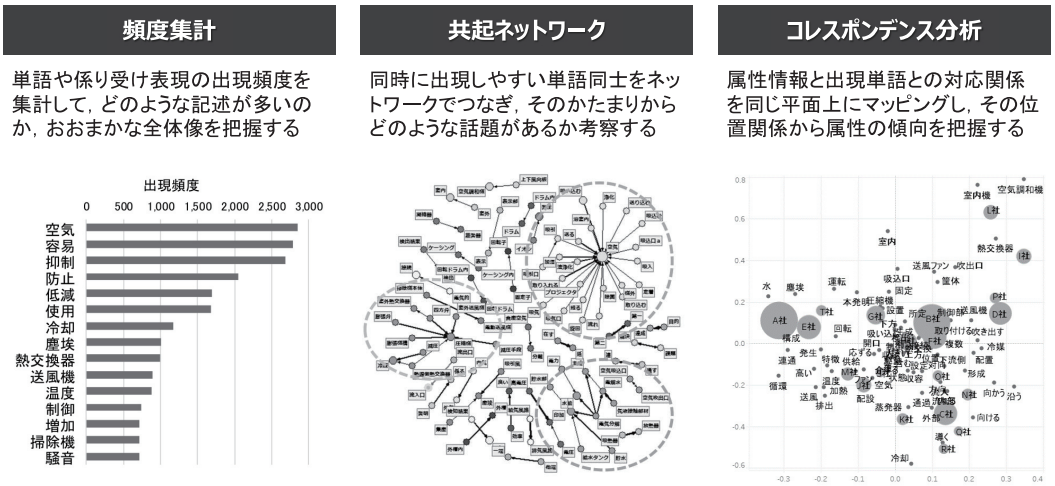


図 3　テキストマイニングのアウトプットの例

1.3　テキストマイニングの複雑性の問題

このようにテキストマイニングは，テキストデータの文書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単位に集計したり，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，現状を把握できる手法となる。

しかし，対象となるテキストデータがビッグデータになると結果がとても複雑になるという問題が生じる。大量のテキストデータからマイニングで抽出される単語は膨大となるため，その大量の単語を単位に可視化される結果は非常に複雑で解釈性が薄れてしまう。これではビジネスの課題解決における有用なインサイトを得ることが困難となる。

この問題の従来の解決策として，抽出される単語を大幅にフィルタリングする試みがあるが，これでは頻出する単語しか分析対象とできず，新たな気づきとなるような細かい個別性のある特徴を把握することが難しい。また，抽出された大量の単語を人が意味的にグルーピングしていくつかのカテゴリを作成し，そのカテゴリを単位にシンプルに可視化をする試みもしばしば見られるが，そのグルーピングルールは属人的となってしまい，さらにその作業負荷はとてつもなく大きい。

ある程度の件数のテキストデータであれば，従来のテキストマイニングのアプローチだけで可視化をしても問題なく傾向を把握できるが，大量のテキストデータとなると従来のアプローチでは限界が生じてしまう。そこで教師なし学習によって使われ方の似ている単語群を機械的に集約できるトピックモデルという技術を適用することが有効となる。

2.　テキストマイニング×トピックモデル

ここではテキストマイニングの複雑性の問題を解決する手段として，トピックモデルを適用することの有効性について述べ，トピックモデルの各手法の概要を解説する。その中でも PLSA という手法の有効性を考察し，PLSA の理論の詳細とテキストマイニングへの適用で得られるアウトプットについて説明する。

2.1　トピックモデルの適用

トピックモデルは主に 1990 年代から 2000 年代初頭にかけて開発が盛んになった自然言語処理技術である。Bag-of-Words による「文書×単語」の行列データをインプットに，教師なし学習で文書に潜むトピックをアウトプットする。そのトピックは文書の集合と単語の集合で構成され，使われ方の似ている単語群は同じトピックに集約

されるようになる。トピックモデルを適用することで，単語単位ではなく集約されたトピックを単位に文書全体の傾向を把握することができる。

教師なし学習でデータを機械的に集約する方法としては，同じく Bag-of-Words のデータに対して，階層クラスター分析の Ward 法や非階層クラスター分析の k-means 法といった従来のクラスター分析を適用する方法も考えられる。しかし，以下の点で挙げられるように，トピックモデルは従来のクラスター分析より Bag-of-Words のデータクラスタリングという観点において優れている。

(1) 高次元データに対応できる

従来のクラスター分析では，データ間の距離に基づいて類似度を計算し，距離の近いデータをまとめていく。しかし，変数の数が大量にある高次元データになると，データ間の距離がどれも大きく離れ，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。Bag-of-Words は大量の単語を列に取った高次元のデータとなるため，こうした従来のクラスター分析手法は適していない。これに対してトピックモデルでは，距離に基づかず，高次元の情報をできるだけ保持した形でその次元数を削減し，低次元に変換できる手法である。そのためトピックモデルはしばしばデータの次元削減でも用いられ，次元削減法と呼ばれることもある。トピックモデルを適用すれば，膨大な単語で構成される Bag-of-Words の高次元データでもトピックに集約することができる。

(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ー分析手法は，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。トピックモデルでは，トピックは行の要素と列の要素の 2 つの軸に基づいて抽出され，各トピックに対する行の要素と列の要素のそれぞれの関連度が同時に計算される。つまり，Bag-of-Words による「文書×単語」の行列データに適用すれば，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解釈できる。このように，抽出されるトピックは行と列の 2 つの軸の情報を持つことができるため，従来のクラスター分析よりも情報量が多い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ー分析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，トピックモデルはソフトクラスタリングとも呼ばれ，すべての要素がすべてのトピックにまたがって所属し，その関連度合いが計算される。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を実現できる。

2.2 トピックモデルの各手法

トピックモデルの代表的な手法には，「LSA」「NMF」「PLSA」「LDA」などがある。ここでは各手法の概要を解説する。

2.2.1 LSA

LSA（Latent Semantic Analysis，潜在意味解析）は，Bag-of-Words による「文書×単語」の行列を SVD（Singular Value Decomposition, 特異値分解）によって次元削減することでトピックを抽出する手法である。1990 年に発表された手法であるが [5]，その基礎となる考え方の研究はそれ以前から存在している。情報検索の分野では LSI（Latent Semantic Indexing）とも呼ばれている。

LSA による特異値分解のイメージを図 4 に示す。LSA の具体的な処理は，「文書×単語」行列を①「文書×トピック」（左特異ベクトル），②「トピック×トピック」（特異値），③「トピック×単語」（右特異ベクトル）の 3 つの行列に分解することである。特異値「トピック×トピック」は対角行列であり，その対角成分には特異値が大きい順に配列される。左特異ベクトルの「文書×トピック」と右特異ベクトルの「トピック×単語」は，トピックを示すベクトル（左特異ベクトルでは各列，右特異ベクトルでは各行）が直交しており，各トピックの軸は数学的に互いに独立している。SVD による数学的な処理は，元の行列を最もよく近似する行列に分解すること，つまり元の行列と分解後の行列の誤差を最小化する最適化問題を解くことである。特に，SVD はこの最適解の存在が数学的に保証されているという点で強力であり，計算過程のシンプルさから計算効率も高いメリットがある。

SVD に類似する手法に PCA（Principal Component Analysis, 主成分分析）があり，PCA は SVD を用いた分析手法と捉えることができるが，両者の関係について簡単に説明する。PCA は元のデータセットの分散を最大化させる方向（主成分）を見つけ出し，最も情報量の多い特徴を抽出することを目指している。その主成分の方向に元のデータを射影することで，元のデータの重要な情報を保持しながら次元を削減する。SVD は分散を最大化させるという目的で用いられるものではないが，PCA を効率的に実行できる手法として用いられるため，PCA は特定の条件下における SVD の応用の一例と見ることができる。つまり，SVD は PCA よりも一般的な手法といえるが，実用面でいえば両者は同様の手法のように扱われることがある。

SVD は大きな値に引っ張られてトピックが抽出される傾向があるため，LSA を実行する際には，Bag-of-Words に前処理として TF-IDF などで重み付けした行列を用いることが多い。また LSA の課題として，分解する行列の要素に負の値を許容しているため，結果の解釈が難しくなることや，結果が学習データに完全に依存するため，過学習を起こしやすく，新しい文書のトピックは推定できないことなどが挙げられる。

LSAの特異値分解

$$X = U\Sigma V^t$$

X	=	U	×	Σ	×	V^t
文書×単語 m×n		① m×k		② k×k		③ k×n

LSAでアウトプットされる各行列の例

①文書×トピック

U	トピック1	トピック2	トピック3	トピック4
文書1	-0.47	-0.75	-0.46	0.11
文書2	-0.61	-0.10	0.69	-0.37
文書3	-0.27	0.41	-0.54	-0.68
文書4	-0.58	0.52	-0.10	0.62

②トピック×トピック

Σ	トピック1	トピック2	トピック3	トピック4
トピック1	7.7	0	0	0
トピック2	0	2.8	0	0
トピック3	0	0	2.0	0
トピック4	0	0	0	0.6

③トピック×単語

V^t	単語1	単語2	単語3	単語4	単語5
トピック1	-0.51	-0.49	-0.38	-0.37	-0.47
トピック2	-0.27	0.14	0.80	-0.50	-0.09
トピック3	0.47	0.43	-0.35	-0.67	-0.14
トピック4	0.61	-0.74	0.23	-0.18	0.07

図 4　LSA の特異値分解

2.2.2　NMF

NMF（Non-negative Matrix Factorization，非負行列因子分解）は，Bag-of-Words の「文書×単語」の行列を，2 つの非負の行列「文書×トピック」と「トピック×単語」に分解することでトピックを抽出する手法で，1999 年に発表された [6]。NMF による行列分解のイメージを図 5 に示す。

計算アルゴリズムは，元の行列と分解後の行列の積との誤差を最小化することを目的関数に，初期値を与えた反復計算により最適解を得る。誤差の定義の仕方は，平方ユークリッド距離や KL ダイバージェンスなどが使われる。反復計算は乗法更新式（Multiplicative update rule）が代表的であり，これは得られる行列が必ず非負であるという制約条件下で適用される勾配降下法である。LSA と比較して，分解後の行列の要素がすべて非負であるため，結果の解釈がしやすいメリットがある。

LSA と異なり，分解された「文書×トピック」と「トピック×単語」の行列は，各トピックのベクトル間で直交しておらず，各トピックは数学的に互いに独立していない。これにより，トピック間に意味的な重複が生じる可能性がある。また，LSA と同様に，NMF でも結果が学習データに完全に依存するため，過学習を起こしやすく，新しい文書のトピックは推定できないという課題がある。

NMFの行列分解

$X = WH^t$

文書×単語 m×n　　① m×k　　② k×n

NMFでアウトプットされる各行列の例

①文書×トピック

W	トピック1	トピック2	トピック3	トピック4
文書1	0.22	0.48	0.16	0.39
文書2	0.35	0.12	0.07	0.88
文書3	0.11	0.58	0.24	0.14
文書4	0.29	0.27	0.60	0.20

②トピック×単語

H^t	単語1	単語2	単語3	単語4	単語5
トピック1	4.34	0.03	1.21	2.24	1.87
トピック2	3.41	1.89	2.81	0.46	1.03
トピック3	0.38	2.24	1.09	5.62	0.06
トピック4	1.85	3.66	0.08	2.11	4.16

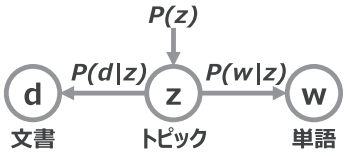
図 5　NMF の行列分解

2.2.3　PLSA

PLSA（Probabilistic Latent Semantic Analysis，確率的潜在意味解析）は，確率モデルを導入してトピックを抽出する手法である。特異値分解によってトピックを抽出する LSA の処理を，確率的な枠組みによって発展させたもので，1999 年に発表された [7]。情報検索の分野では PLSI（Probabilistic Latent Semantic Indexing）と呼ばれることもある。

PLSA では，Bag-of-Words の「文書×単語」の行列を確率的に分解することでトピックを抽出する。PLSA の確率モデルのイメージを図 6 に示す。PLSA のモデルは文書 d と単語 w の同時確率 P（d, w）を，3 つの確率分布① P（d|z），② P（w|z），③ P（z）に分解して表現する。この 3 つの確率分布を Bag-of-Words の「文書×単語」行列データに基づき推定する。具体的には，P（d, w）の対数尤度関数を最大化する EM アルゴリズムを実行し，初期値を与えた反復計算により最適解を得る。

PLSAの確率モデル



$$P(d,w) = \Sigma_z P(d|z)P(w|z)P(z)$$

PLSAの出力結果例

P(文書d | トピックz)

P(d|z)	トピック1	トピック2	トピック3	トピック4
文書1	0.41	0.16	0.06	0.27
文書2	0.29	0.54	0.11	0.06
文書3	0.22	0.13	0.04	0.58
文書4	0.08	0.17	0.79	0.09

$\Sigma_d P(d|z) = 1$

P(単語w | トピックz)

P(w|z)	トピック1	トピック2	トピック3	トピック4
単語1	0.11	0.09	0.44	0.20
単語2	0.09	0.38	0.04	0.18
単語3	0.50	0.11	0.29	0.09
単語4	0.08	0.14	0.17	0.31
単語5	0.22	0.28	0.06	0.22

$\Sigma_w P(w|z) = 1$

P(トピックz)

P(z)	トピック1	トピック2	トピック3	トピック4
	0.31	0.27	0.23	0.19

$\Sigma_z P(z) = 1$

図 6　PLSA の確率モデル

PLSA のメリットとして，LSA では「文書×単語」の行列に対して事前に TF-IDF などで重みづけする必要があったが，PLSA では確率的な処理によりそうした重みづけをせずに実行できる。また，PLSA は推定される P(d|z) と P(w|z) によって，文書および単語のトピックに対する関連度が所属確率として出力されるため，結果が解釈しやすいメリットがある。なお，各トピックは互いに独立しているという数学的な仮定を置いている。

PLSA のデメリットとしては，入力する「文書×単語」のデータセットが大規模になると計算コストが高くなることが挙げられる。また，PLSA も LSA や NMF と同様に，結果が観測データに完全に依存するため，過学習を起こしやすく，新しい文書のトピックは推定できないという課題があるが，一方で，観測データに対する再現度は高く，その観測データ固有の特徴をそのまま反映できるモデルと捉えることもできる。

2.2.4 LDA

LDA（Latent Dirichlet Allocation，潜在ディリクレ配分法）は，PLSA をベイズ的に拡張させた手法で，ディレクレ分布を事前分布に導入しており，2003 年に発表された [8]。

LDA のグラフィカルモデルを図 7 に示す。LDA は確率的な生成モデルとも言われ，文書 d 内の各単語 w が特定のトピック z から生成されたと仮定する。このトピック z の分布は文書 d ごとに異なり，その確率分布 P（z|d）は θ で表される。一方，特定のトピック z が各単語 w を生成する確率分布 P（w|z）は ϕ で表し，そのトピック z の下で単語 w を生成するプロセスでは ϕ が参照される。つまり，LDA では $\theta \rightarrow z \rightarrow w$ という生成プロセスが仮定され，その中で $z \rightarrow w$ というアローの確率過程は $\phi \rightarrow w$ というアローによる確率分布 ϕ が参照される。なお，θ と ϕ はそれぞれハイパーパラメータ α と β を持つディレクレ分布に従う。LDA の推定対象は θ = P（z|d）と ϕ = P（w|z）であるが，これらの生成過程を逆にたどるアルゴリズムにより，単語 w（観測データ）から θ と ϕ を推定する。具体的には，各文書 d 中の各単語 w（観測データ）に対して，まずその単語 w がどのトピック z から生成されたと考えられるか，θ と ϕ の暫定値（初期値）から推定し（z の推定），そしてその結果を基にして，各文書 d がどのトピック z をどれだけ含むか推定し（θ の推定），各トピック z がどの単語 w を生成する可能性が高いか推定する（ϕ の推定）。これを繰り返すことで θ と ϕ を逐次的に更新していく。推定アルゴリズムは，ギブスサンプリングや変分ベイズ法などが適用され，反復的に計算される。

PLSA はパラメータが固定的で，観測データのみから直接推定するため，結果が完全に観測データに依存するが，LDA はパラメータが事前分布に従って変動する確率分布とし，観測データと事前分布から推定する。事前分布を導入することで，確率的なスムージング効果があり，過学習を抑制できる。また，LDA は新しい文書についても推定ができ，新しい文書に対応する θ を未知とし，新しい文書に含まれる単語 w（観測データ）と学習済みの ϕ と α から θ を推定する。

一方，LDA はハイパーパラメータ α と β によって結果が変動しやすく，この値の設定の仕方，推定の仕方が難しい。ハイパーパラメータ α, β の大きさとディレクレ分布の特徴は，$\alpha, \beta > 1$ では，値が大きいほど確率が均一化し，どの文書も同じようなトピックの分布を持つようになり，トピックは多様な単語に関連づけられる傾向がある。$\alpha, \beta = 1$ では分布は一様分布となり，様々なトピックにランダムに確率が割り当てられる。これは事前分布を仮定していない PLSA と近い振る舞いをする。$1 > \alpha, \beta > 0$ では，値が小さいほど確率が局所的になり，文書は特定のトピッ

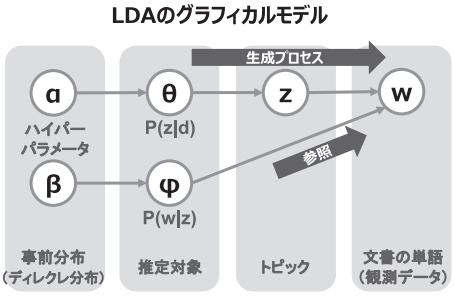


図 7　LDA のグラフィカルモデル

クに確率が集中し，トピックは一部の単語に偏って関連づけられる傾向がある。また，LDA は新しい文書の推定ができる高い汎化性能を持つ一方で，トピックの結果が一般的で抽象度が高くなることがある。

なお，PLSA と LDA はどちらも文書 d と単語 w の関係をモデル化している手法だが，文書 d と単語 w の関係の扱い方の違いには注意が必要である。PLSA の推定対象は P（d|z）と P（w|z）であるが,「文書 d ×単語 w」の行列データに対して文書 d と単語 w の役割には対称性があり，行と列を入れ替えても推定結果に影響はない。一方，LDA の推定対象は P（z|d）と P（w|z）であるが，文書 d と単語 w の役割は対称的ではなく，それぞれの役割は明確に区別されたモデルとなっている。そのため，行と列を入れ替えて適用すれば異なる結果となる。

2.3　テキストマイニングに適用するトピックモデルの考察

トピックモデルの中では LDA を適用することが一般的である。しかし，LDA があらゆる課題で優れたトピックモデルということではない。トピックモデルの各手法の特徴を俯瞰的に理解し，解決したい課題に応じて使い分けることが重要である。

LSA，NMF，PLSA は，過学習を起こしやすいことがデメリットとして挙げられ，そうした背景から LDA が開発されているが，これは新しい文書の推定を課題としたときのデメリットである。たとえば，情報検索などへの適用を想定した場合，学習データにないテキストデータのトピックを推定できることは確かに重要であり，LDA を適用することが相応しいと考えられる。では，テキストマイニングに適用する場合はどうか考察する。先述したテキストマイニングの複雑性の問題を解決する手段としてトピックモデルを適用する場合，テキストマイニングが現状把握をするための手法であることを考えれば，新しいデータのトピックを推定することよりも，観測データのありのままの特徴をトピックで反映できることが重要といえる。LSA，NMF，PLSA は観測データに依存した過学習でトピックを抽出するモデルだが，観測データに対する再現度は高く，その観測データ固有の特徴をそのまま反映できるモデルと捉えることができる。

この点において，LDA は事前分布にディレクレ分布を仮定していることで，観測データをそのまま再現するモデルではなくなっている。また LDA は新しい文書の推定ができる高い汎化性能を持つ一方で，抽出されるトピックの内容が一般的で抽象度が高くなることがある。さらに，LDA はディレクレ分布のハイパーパラメータによって結果が変動しやすく，その設定の仕方，推定の仕方が難しい。テキストマイニングは専門知識がなくても直感的に操作できる使いやすさが特長の一つであるため，ビジネス現場では文系理系問わず人気の分析手法となっている。こうした点においても，より高度な LDA を適用することは複雑で扱いにくさが生じてしまう懸念がある。

したがって，テキストマイニングの複雑性の問題を解決する手段としては，LDA よりも LSA，NMF，PLSA の方が適していると考えられる。その中でも PLSA は，LSA のような TF-IDF などの前処理がいらず，トピックの結果が確率で解釈しやすく，かつ各トピックが独立しているという特徴を有していることから，筆者は PLSA がテキストマイニングに適用するトピックモデルの手法として相応しいと考えている。

2.4　PLSA の適用によるトピック抽出

ここでは PLSA という手法についてより詳細に解説する。PLSA は自然言語処理に限らずデータクラスタリングや次元削減法として用いられる手法である。改めて PLSA によるクラスタリングのイメージを図 8 に示す。自然言語処理に限定せず，より一般的に PLSA という手法を表現すると，共起行列と呼ばれる行列データをインプットに教師なし学習をし，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 を抽出する手法となる。自然言語処理の分野では，共起行列を Bag-of-Words の「文書×単語」行列とし，潜在クラスをトピックと呼んでいる。つまり，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，文書と単語に共通するトピックを抽出している。これにより文書全体をいくつかのトピックを軸に解釈できるとともに，使われ方の似ている単語群をそのトピックの構成要素として集約できる。自然言語処理の分野以外では，たとえば「画像×画像特徴量」の行列データ（Bag-of-visual-words と呼ばれるデータ）に PLSA を適用して画像分類を行う取り組みや [9]，「顧客 ID ×購買商品」の行列データ（ID-POS データと呼ばれるデータ）に PLSA を適用して顧客と商品を同時に分類する取

り組みなどがある[10]。

以下にPLSAによるトピック抽出までの計算過程を説明する。PLSAでは文書xにおける単語yの出現確率P(y|x)を，潜在的なトピックzを介した式（1）でモデル化する。これはAspectモデルと呼ばれるが，Aspectとは文書内に潜むトピックを指している。このAspectモデルに，確率の連鎖律により成り立つ同時確率と条件付確率の関係式(2)と，ベイズの定理により成り立つ式(3)を適用する。これによって式(4)で示すように，文書xと単語yの同時確率P(x, y)を３つの確率分布①P（x|z），②P（y|z），③P（z）に分解して表現する。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（5）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計算する。つまり式（6）のEステップと式（7）～（9）のMステップを計算することで最尤推定する。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|z），P（y|z），P（z）の値となる。

図２で示したホテルのレビューデータのBag-of-WordsにPLSAを適用した場合，アウトプットとして得られる「文書×トピックの確率分布P（x|z)」「単語×トピックの確率分布P（y|z)」「トピックの周辺確率P（z)」のイメージを図９に示す。図９はホテルのレビューデータから「広さ」「清潔さ」「利便性」「スタッフ」「朝食」という５つのトピックが抽出されたと仮定したときの結果のイメージとなる。P（x|z）はトピックzに対する各文書xの所属確率を示しており，P（y|z）はトピックzに対する各単語の所属確率を示している。P（x|z）はそれぞれのトピックにおいてすべての文書の所属確率を合計すると100％となり，P（y|z）はそれぞれのトピックにおいてすべての単語の所属確率を合計すると100％となる。またトピックの周辺確率であるP（z）も，すべてのトピックの確率を合計すると100％となる。こうした確率分布として得られるアウトプットは解釈性に優れており，PLSAを適用することで，どのような単語群で構成されるトピックが抽出され，それぞれの文書はどのトピックの要素が多いのか把握できる。

このようにテキストマイニング×PLSAの適用により，大量のテキストデータを複雑な単語単位ではなく，集約されたトピックを単位としてシンプルに傾向を分析できる。

$$\mathrm{P}(y \mid x) = \sum_z \mathrm{P}(y \mid z)\mathrm{P}(z \mid x) \tag{1}$$

$$\mathrm{P}(x, y) = \mathrm{P}(y \mid x)\mathrm{P}(x) \tag{2}$$

$$\mathrm{P}(z \mid x) = \frac{P(x \mid z)P(z)}{P(x)} \tag{3}$$

$$\mathrm{P}(x, y) = \sum_z \mathrm{P}(x \mid z)\mathrm{P}(y \mid z)\mathrm{P}(z) \tag{4}$$

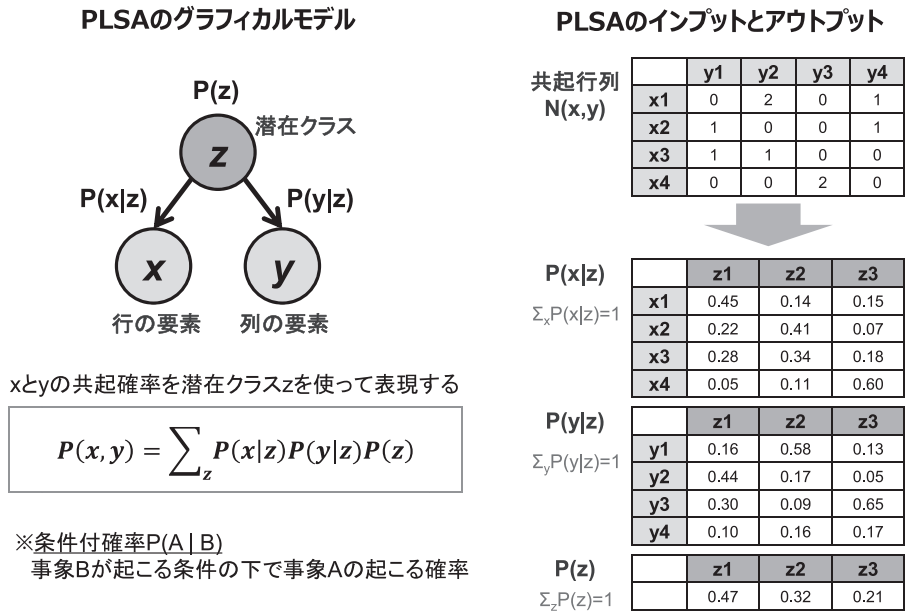
$$L = \sum_x \sum_y \mathrm{N}(x, y) \log \mathrm{P}(x, y) \tag{5}$$

$$\mathrm{P}(z \mid x, y) = \frac{\mathrm{P}(x \mid z)\mathrm{P}(y \mid z)\mathrm{P}(z)}{\sum_z \mathrm{P}(x \mid z)\mathrm{P}(y \mid z)\mathrm{P}(z)} \tag{6}$$

$$\mathrm{P}(x \mid z) = \frac{\sum_y \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)}{\sum_x \sum_y \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)} \tag{7}$$

$$\mathrm{P}(y \mid z) = \frac{\sum_x \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)}{\sum_x \sum_y \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)} \tag{8}$$

$$\mathrm{P}(z) = \frac{\sum_x \sum_y \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)}{\sum_x \sum_y \sum_z \mathrm{N}(x, y)\mathrm{P}(z \mid x, y)} \tag{9}$$



共起行列 N(x,y)	y1	y2	y3	y4
x1	0	2	0	1
x2	1	0	0	1
x3	1	1	0	0
x4	0	0	2	0

P(x|z)	z1	z2	z3
x1	0.45	0.14	0.15
x2	0.22	0.41	0.07
x3	0.28	0.34	0.18
x4	0.05	0.11	0.60

P(y|z)	z1	z2	z3
y1	0.16	0.58	0.13
y2	0.44	0.17	0.05
y3	0.30	0.09	0.65
y4	0.10	0.16	0.17

P(z)	z1	z2	z3
	0.47	0.32	0.21

図 8　自然言語処理に限定しない PLSA によるクラスタリング

文書×トピックの確率分布 P(x|z)

文書ID	ホテルの口コミコメント	広さ	清潔さ	利便性	スタッフ	朝食
1	部屋が広く快適で、部屋の空調も良く効きました。	**0.07 %**	0.04 %	0.01 %	0.02 %	0.01 %
2	部屋は綺麗に清掃されていて、お風呂も快適でした。	0.04 %	**0.09 %**	0.01 %	0.03 %	0.01 %
3	駅の近くで便利で良く、部屋も広くて良かったです。	**0.06 %**	0.01 %	**0.08 %**	0.02 %	0.02 %
4	スタッフの対応が良く、部屋も丁寧に清掃されていました。	0.03 %	**0.07 %**	0.01 %	**0.08 %**	0.02 %
5	朝食が美味しく、レストランも広くて綺麗で良かったです。	0.05 %	0.04 %	0.01 %	0.02 %	**0.09 %**
…	……					
	合計	100 %	100 %	100 %	100 %	100 %

トピックの確率分布 P(z)

トピック	P(z)
広さ	26 %
清潔さ	22 %
利便性	20 %
スタッフ	18 %
朝食	14 %
合計	100%

単語×トピックの確率分布 P(y|z)

トピック	部屋	広い	快適	空調	良い	効く	綺麗	清掃	風呂	駅	近い	便利	スタッフ	対応	丁寧	朝食	美味しい	レストラン	…	合計
広さ	**12%**	**18%**	**14%**	0.9%	4%	0.5%	5%	3%	8%	0.3%	0.1%	0.2%	0.6%	0.7%	0.4%	0.8%	0.3%	2%		100%
清潔さ	9%	2%	3%	1%	3%	0.4%	**19%**	**15%**	6%	0.4%	0.2%	0.1%	4%	2%	3%	0.9%	0.4%	1%		100%
利便性	0.8%	0.4%	2%	0.1%	7%	0.3%	0.2%	0.3%	0.5%	**11%**	**17%**	**20%**	1%	0.8%	1%	0.6%	1%	6%		100%
スタッフ	2%	0.9%	1%	0.6%	5%	0.3%	1%	3%	1%	0.6%	0.3%	0.2%	**16%**	**21%**	**15%**	0.7%	0.2%	1%		100%
朝食	1%	1%	0.7%	0.8%	3%	0.6%	1%	0.8%	0.3%	0.9%	1%	0.7%	2%	1%	0.8%	**23%**	**17%**	**14%**		100%

図 9 「文書×単語」行列に PLSA を適用して得られるトピックの確率分布の例

3.　テキストマイニング×ベイジアンネットワーク

ここではテキストマイニングにベイジアンネットワークという確率モデルの AI 技術を応用することの有用性について述べる。テキストマイニングの分析が表層的な傾向の現状把握に留まっていることを指摘した上で，それを解決する手段として，データに潜む要因関係をモデル化し，確率的な推論を実行できるベイジアンネットワークの概要と応用可能性を説明する。

3.1　テキストマイニングの現状把握分析の限界

テキストマイニングの分析では，全体像としてどのような単語が多く使われている傾向にあるのか現状を把握できる。また，アンケートデータにおける回答者の性別・年代や，特許文書データにおける出願年や出願人といった属性情報がテキストデータに紐づいていれば，その属性別に出現しやすい単語の傾向の違いも把握できる。さらにテキストマイニングにトピックモデルを適用すれば，この傾向をトピックという分かりやすい単位で把握できる。

ただ，この傾向分析で把握できるのは表層的な傾向であり，その属性で出現しやすい単語あるいはトピックが多い順に列挙された結果となる。その属性との統計的な関係の強さの可視化や，その出現傾向に影響を与えている要因の探索まではできていない。また，テキストマイニングは現状を把握する分析であり，その現状から状況が変化したときに，その変化に伴って結果がどのような影響を受けるのかシミュレーションすることはできない。ビジネスの課題解決において，まずは現状を知ることはきわめて重要であるが，知るだけでは課題解決にはならない。その現状がどのような要因と関係しており，どのようなアクションを取れば現状の結果をどの程度コントロールできるのか，こうした要因構造まで探索することが課題解決の施策検討において要となる。

このようにテキストデータに潜む要因の関係性まで探索するような深い分析を実行する手段として，ベイジアンネットワークという確率モデルの AI 技術を適用することが有効となる。

3.2　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る[11)]。このモデルを用いることで，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。ベイジアンネットワークの概要図を図 10 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すべ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の連鎖律で表現している。そのリンク構造は，観測データと変数の定義に基づき，それを数学的に最もよく説明する（尤度を最大化する）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。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

どの変数がどの変数に影響しているのか，構築されたモデルの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。なお，ベイジアンネットワークでモデル化される有向リンクの関係性は，本当の因果関係ではなく，あくまでも確率的な因果関係である。本当の因果関係は原因と結果の関係であるが，確率的な因果関係では，その変数を条件に与えた条件付確率を形成する方がモデルの尤度（データへの当てはまりの良さ）が向上するというものである。

(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。たと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，それ以外の部分を観測データから数学的に探索するということもできる。たと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) 互いの変数間の関係を把握しさまざま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，さまざま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であるため，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。

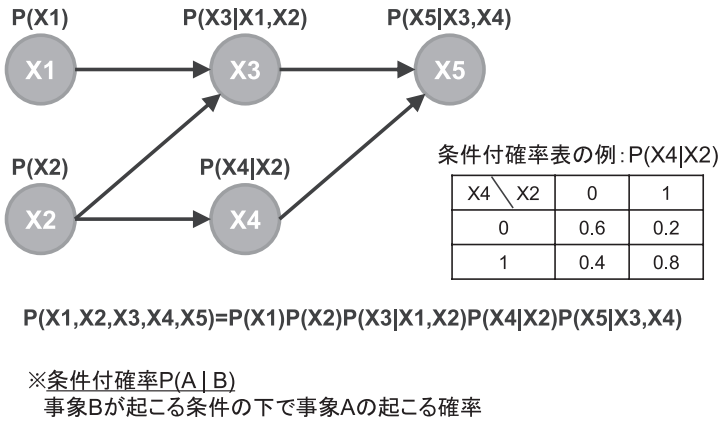


図 10　ベイジアンネットワークの概要図

3.3　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

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。たとえば，テキストマイニングによって構築された Bag-of-Words のデータの単語一つひと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 [12)]。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 4,238 件を対象に，事故状況がテキストで記されたデータにテキストマイニングを実行することで事故に関わる製品や行動の単語を抽出し，それら一つひとつを確率変数としている。また子どもの性別・年齢の情報や事故の情報も確率変数とし，これらの関係を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 11 にて紹介する。このベイジアンネットワークのモデルを用いることで，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，どのような事故にいた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。

このように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メリットをまとめると，主に以下の 2 つが挙げられる。

(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，文書でどのような単語が多く使われる傾向があるのか，またその傾向が属性によってど

のような違いがあるのか把握できる。ここにベイジアンネットワークをさらに適用することで，そうした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，統計的な関係性を構造的なモデルとして可視化できる。たとえば図 11 のモデルでは，子どもの製品に対する行動はどのような発達段階を要因としており，またどのような事故の要因となり得るのか，その統計的な関係性を分析できている。また，図 2 ではホテルのレビューデータの Bag-of-Words を例に取り上げているが，レビューデータには投稿者の性別や年代，評点といった属性情報が紐づいていることが多い。ベイジアンネットワークを適用すれば，こうした属性情報とレビューコメントの内容の間に存在する統計的な関係性もモデル化できる。

(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

先述したとおり，テキストマイニングは基本的に文書の記述傾向の現状把握をする手法であるが，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・影響を受けるのかシミュレーションすることはしない。ここにベイジアンネットワークをさらに適用することで，その確率的なシミュレーションが可能になる。上記 (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，テキストデータに潜む要因関係を構造的にモデル化できるようになる。そのモデルを利用することで，たとえば図 11 のモデルでは，子どもの発達段階の変化によって行動の確率がどの程度変化するのか，同じ行動でも製品が異なれば起こり得る事故の確率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することができている。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計算されている。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 歳児で確率がさらに高くなることがシミュレーションで計算されている。こうした分析によって，子どもの事故の発生を抑制するための具体的な施策をデータドリブンで検討できる。

また，図 2 で示したようなホテルのレビューデータにベイジアンネットワークを適用し，レビューに紐づく投稿者の性別・年代や評点という属性を確率変数に組み込んだモデルを構築すれば，そのモデルを用いて以下のようなシミュレーションが可能である。たとえば性別・年代が変わるとレビュー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レビューの評点がどの程度変化するのか，確率的にシミュレーションを実行できる。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る。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することや，その効果を最大化させる条件を探索することもでき，ビジネスの課題解決において効果的なアクションを検討することができる。

3.4 テキストマイニング×ベイジアンネットワークの課題

図 11 のモデルで取り上げた分析例では，テキストマイニングで抽出された単語一つひと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。つまり，Bag-of-Words のデータにそのままベイジアンネットワークを適用して，単語一つひとつを確率変数にすると重要な傾向や気づきが埋もれてしまう懸念がある。本来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り，情報が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。

そこでテキストマイニングに直接ベイジアンネットワークを適用するのではなく，まずトピックモデルを適用することを考える。この処理を挟むことによって，単語一つひとつではなく，単語群が集約されたトピックを確率変数として定義し，その上でベイジアンネットワークを適用する。つまり，Bag-of-Words ではなく Bag-of-Topics のデータにベイジアンネットワークを適用する。これによって，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。こ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の 3 つの手法を組み合わせてテキストデータを分析する手法を次項で解説する。

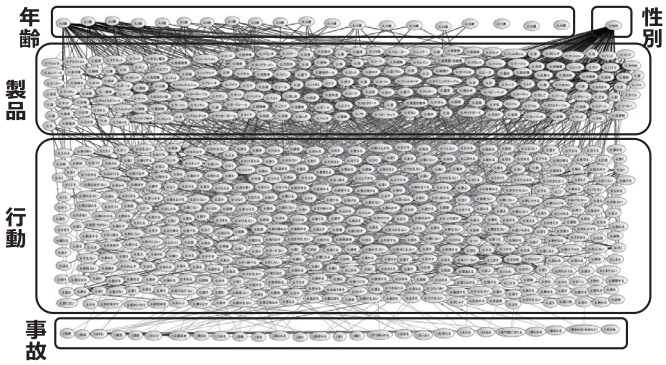


図 11　Bag-of-Words にベイジアンネットワークを適用したモデルの例（※文献 12) より引用）

4.　テキストマイニング× PLSA ×ベイジアンネットワークの分析手法：Nomolytics

ここでは従来のテキストマイニングに加え，トピックモデル（PLSA）とベイジアンネットワークという 2 つの AI 技術を連携して適用し，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）[13] 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）[14]，令和 6 年 4 月 30 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。

4.1　Nomolytics という分析手法

Nomolytics の概要を図 12 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の文書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する。次に，その共起行列をインプットとして PLSA を適用し，使われ方の似ている単語群を複数のトピックにまとめ上げる。最後に，すべてのテキストデータに対して各トピックの該当度を計算することでトピックを確率変数として定義し，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する。

こうした 3 つの手法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さまざま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シミュレーションを実行することができる。

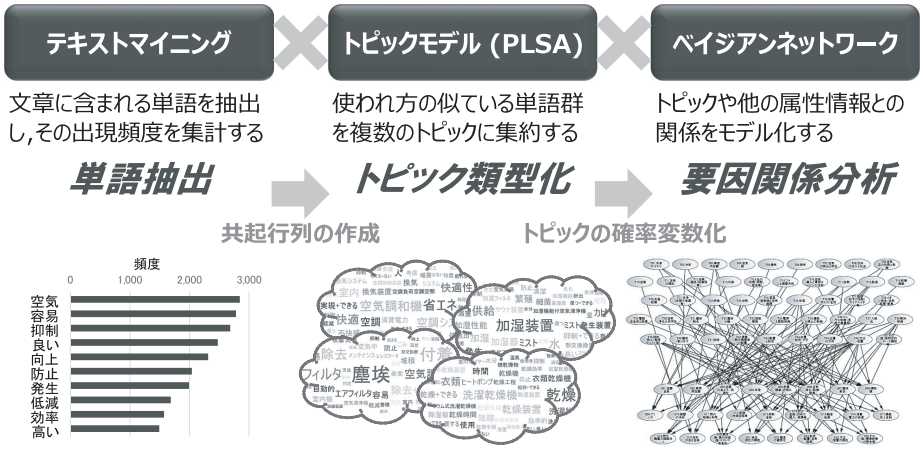


図 12　Nomolytics の概要図

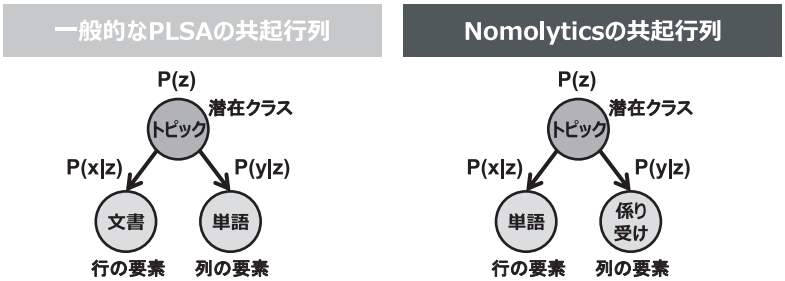
4.2　Nomolytics における各手法の連携の工夫

Nomolytics では，「テキストマイニング× PLSA」と「PLSA ×ベイジアンネットワーク」というそれぞれの手法を連携する際の処理に独自の工夫を施しており，それが Nomolytics の技術的な特徴にもなっている。以下にその連携の工夫について紹介する。

4.2.1　PLSA の共起行列の構成の工夫

先述したトピックモデルの解説では，トピックモデルは従来のクラスター分析と異なり行と列を同時にクラスタリングできると説明した。つまり，PLSA の共起行列の行と列は，双方が十分意味を持つ情報で構成すれば，抽出されたトピックの意味は 2 つの軸から解釈することができる。図 13（左）に示すように，一般的な PLSA の適用では，「文書×単語」という構成の Bag-of-Words を共起行列としてインプットする。しかし，行に設定された「文書」とは文書 ID の情報であり，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 Bag-of-Words は，ほとんどは 0 となる疎なデータであるため，文書間・単語間で違いが現れにくく，特徴がクリアなトピックが得られにくい。

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ている。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 13（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）とする方法を取る。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と，その観点の具体的な内容となる係り受け表現を軸にクラスタリングができ，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る。また，この共起行列は各単語と各係り受けが同時に出現する文章数（共起頻度）が値として入るクロス集計型の行列であり，一般的な PLSA の共起行列のような 0 ばかりの疎なデータとは異なり密なデータとなりやすくなる。これにより，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トピックを抽出でき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いものになり，計算時間も大幅に削減できる効果がある。



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

	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 13　Nomolytics における PLSA の共起行列構成の工夫

4.2.2　トピックの確率変数化の工夫

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，図 9 で示したような所属確率という値によっ

て各文書 ID 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所属確率が計算されているが，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，各テキストデータの文書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，トピックの確率変数を作成する。

4.3　Nomolytics の適用例

Nomolytics はテキストデータであればあらゆる分野で適用できるが，ここではテキストデータの種類ごとにその適用例をいくつか紹介する。具体的には,「自由記述付アンケートデータ」「ユーザレビューデータ」「コールセンターの問い合わせ履歴データ」「特許文書データ」の 4 種類のデータについて，Nomolytics を適用して分析するアプローチの例を紹介する。なお，ここで取り上げる適用例は，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 Nomolytics を適用した事例に基づいている。

4.3.1　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だが，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はそれぞれ独立した分析がされがちである。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ある。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る。一方，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，とても貴重な情報といえる。

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する。そうすることによって，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る。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る。

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 14 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る。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る。またベイジアンネットワークの確率シミュレーションを実行すれば，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る。なお，図 14 のモデル図は，イメージしやすいように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ある。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る。

また，図 15 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関係構造をモデル化することもできる。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る。そしてベイジアンネットワークの確率シミュレーションを実行すれば，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る。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あり，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る。

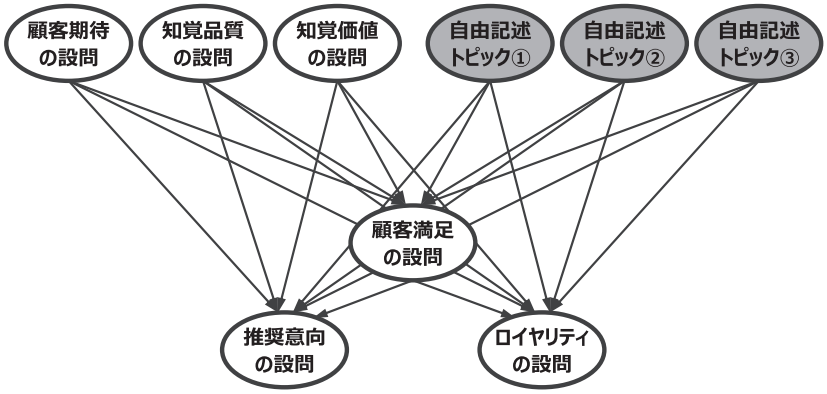


図 14　自由記述アンケートデータを用いた顧客満足の構造モデル

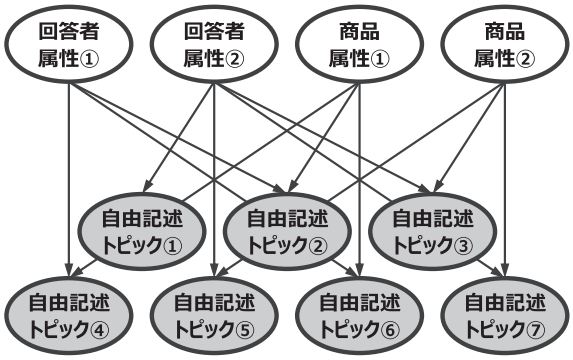


図 15　自由記述アンケートデータを用いた自由記述内容の構造モデル

4.3.2　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，口コミとも呼ばれる。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 上で閲覧できる。

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るが，取得できないものもある。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るが雑なものや，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もあるなど，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ない。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る。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る。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，競合他社の分析などに活用する試みがされている。なお，レビューデータは Web 上で誰でも閲覧できるが，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しなければならない。

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は目視で 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

実施されている。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る。

たとえば，図 16 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る。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トピックから評点に影響を与える要因を把握できる。また，ベイジアンネットワークの確率シミュレーションによって，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る。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る。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，マーケティング戦略などへの活用が考えられる。

また，先述したアンケートデータのモデル化と同様に，図 17 に示すような，レビューのトピックに対する関係構造をモデル化することもできる。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る。そしてベイジアンネットワークの確率シミュレーションによって，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る。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る。

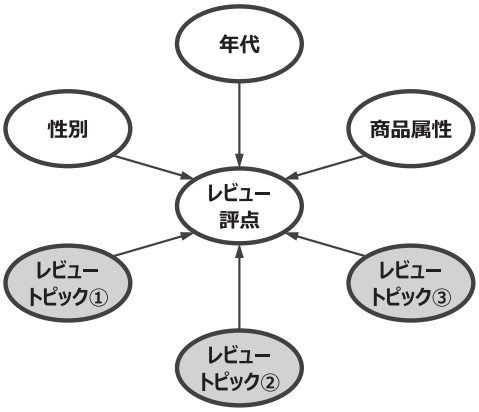


図 16　ユーザレビューデータを用いた評点の構造モデル

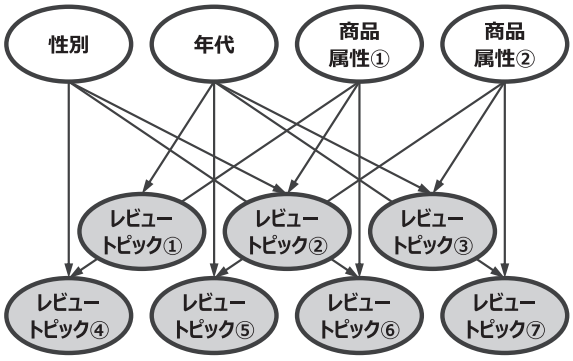


図 17　ユーザレビューデータを用いたレビュー内容の構造モデル

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する。宿泊施設のレビューデータを用いたモデル化の事例では [15]，京都府にある宿泊施設のレビューデータ 11,535 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が，図 18 に示すようにモデル化されている。このモデルを用いたベイジアンネットワークの確率シミュレーションの結

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た。

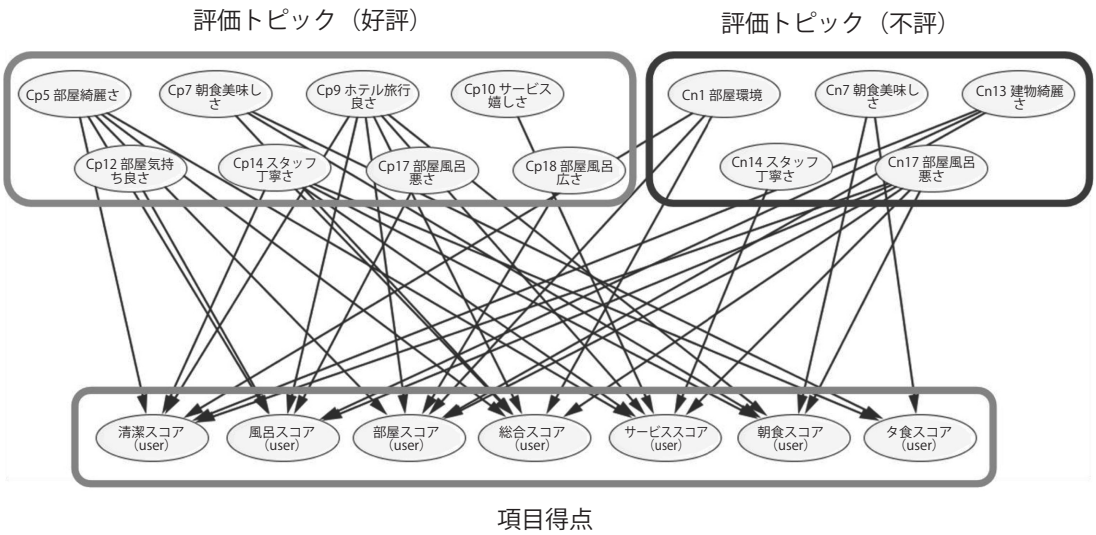


図 18　宿泊施設のレビューデータを用いた評点に対するトピックの関係モデル（※文献 15) より引用）

また，観光地のレビューデータを用いたモデル化の事例では [16]，国内の温泉地におけるレビューデータ 12,564 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）との関係が図 19 に示すようにモデル化されている。このモデルを用いたベイジアンネットワークの確率シミュレーションの結果から，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という結果が得られた。一方，女性はカップル・夫婦とは行かない（友達や家族などと行く）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る。先のシミュレーション結果を活用すれば，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 するといった施策を検討できる。一方，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 するといった施策を検討できる。

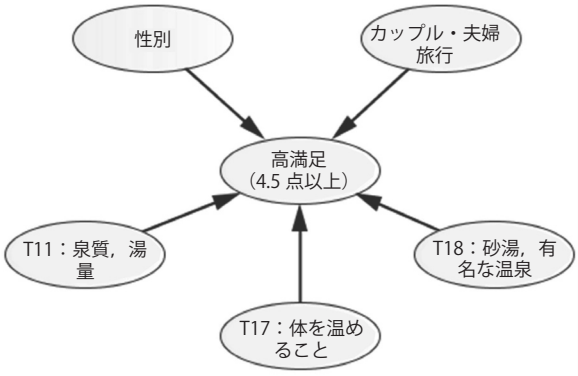


図 19　観光地のレビューデータを用いた評点に対する観光客属性とトピックの関係モデル（※文献 16) より引用）

4.3.3　コールセンターの問い合わ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い合わせ情報は，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る。問い合わせデータの記録対象は企業によってさまざまだが，その問い合わ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い。取得可能な場合は，問い合わせをした顧客の属性情報も紐づけられている。また，問い合わ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い合わ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，問い合わ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る。他に，契約商品の解約要求や会員の退会要求といった，問い合わ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る。

企業のコールセンター部門では，顧客の問い合わせに対応することがメインの業務であり，その問い合わ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る。つまり，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ないが，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ある。こうした顧客の問い合わ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る。実際にコールセンターの問い合わせデータの分析は，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だが，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。しかし，コールセンターの問い合わせデータは，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ある。

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う。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ある。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く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る。さらにそこにベイジアンネットワークを適用すれば，顧客の問い合わ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る。

ここではコールセンターの問い合わせデータに Nomolytics を適用してモデルを構築するアプローチの例を紹介するが，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ある。たとえば，各問い合わ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い合わせの状態に関する情報や，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，問い合わ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。構築されるモデルのイメージを図 20 に示す。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い合わ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る。そしてベイジアンネットワークの確率シミュレーションを実行することで，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い合わ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る。

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 21 に示すような問い合わ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る。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い合わせをする傾向にあるのか，その確率をシミュレーションできる。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ができる。また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い合わ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ため，問い合わせ対応の事前準備などにも活用が考えられる。

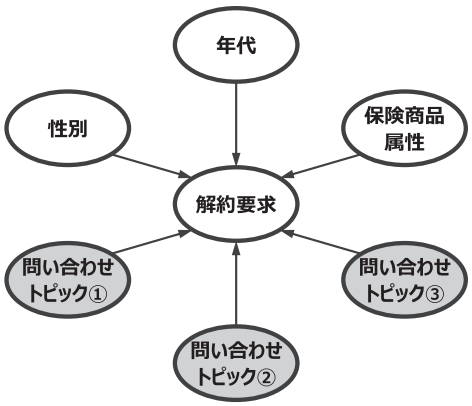


図 20　コールセンター問い合わせデータを用いた顧客の要求行動の構造モデル

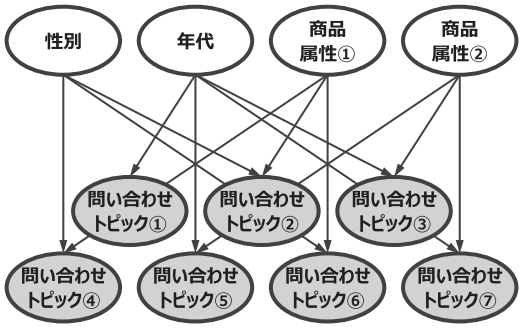


図 21　コールセンター問い合わせデータを用いた問い合わせ内容の構造モデル

4.3.4　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する。

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しばしば分析の対象とされる。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ある。特許情報分析というと従来パテントマップ[17]と呼ばれるものが代表的であり，これは主に特許の出願年や出願人，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。

また特許情報の要約文や請求項，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を適用することで，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取り組みもされている[18]。その分析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る。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 2 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だが），その記述形式を利用して分析するものである。「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，その単語群を分析者がカテゴリに分類し，図 22 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する。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しようとするとき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る。

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い。先述のとお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あり，カテゴリ化のルールも属人的になってしまう。また従来の分析アプローチでは，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るが，統計的な関係の分析は

されていない。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らない。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る。

こうした従来の分析アプローチに対して，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く。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する。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を適用にすることで，図 23 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。これにより，用途と技術の確率統計的な関係性を把握できる。

次項以降では，Nomolytics を特許文書データに適用した具体的な分析事例を紹介する。

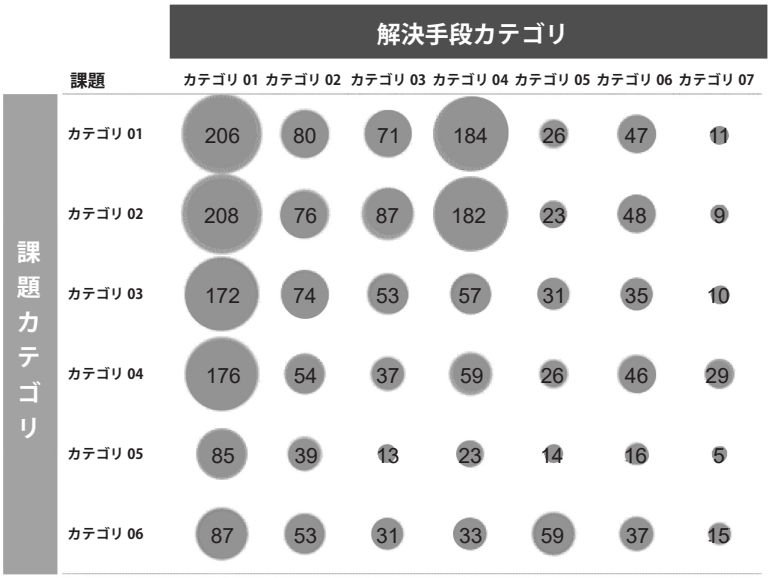


図 22　特許の課題と解決手段をカテゴリ化してクロス集計した従来の分析例

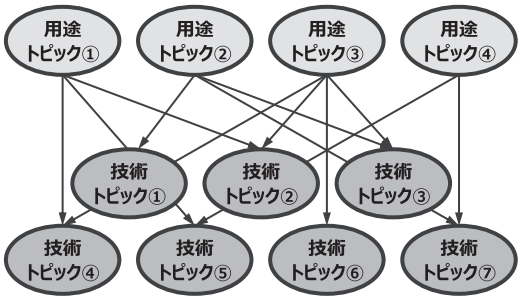


図 23　特許文書データを用いた用途と技術の関係モデル

5.　Nomolytics を特許文書データに適用した分析事例（その 1）

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。なお，本項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い。あくまでも対外紹介用に作成した分析事例となるが，ここで紹介している分析のアプローチ自体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となる。

5.1　分析で用いるデータ

本分析事例では，特許の要約と請求項に「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 件を分析対象としている。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。ここでは特許の要約文の記述情報を主な分析対象としている。国内特許の要約文は 400 字以内という字数制限が設けられており，また厳密なルールではないが，【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る。要約文の例を図 24 に示す。「風」「空気」をキーワードに含む特許のデータということで，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る。またデータの属性情報として，「出願年」や「出願人」などの情報が紐づいている。

【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る

図 24　分析対象とする特許の要約文の例

5.2　分析の全体像

本分析事例では，特許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。さらに，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 25 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とおりとなる。

(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」の属性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」の属性で集計することで，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，自社の技術の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。この関係分析は，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。

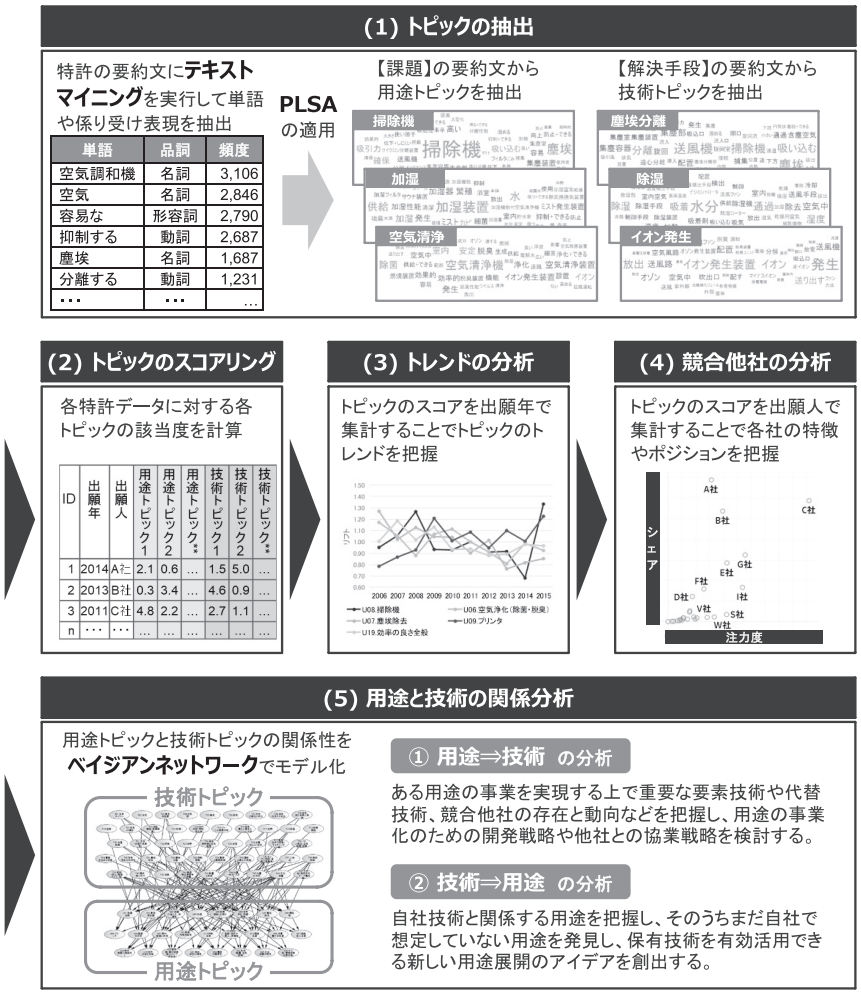


図 25　特許文書データに Nomolytics を適用した分析のプロセス

5.3　PLSA の適用による用途と技術のトピック抽出

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。先述のとおり，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。以下にトピック抽出の手順を述べる。

(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。課題と解決手段という項目のない特許は，要約文全体を対象とする。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およびその逆の係り受け（動詞・形容詞・形容動詞に対する名詞）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)を使用した。本ツールの形態素解析において，文章の区切りを定義する文字列には「句点（。)」の他に，特許要約文の中で文章のラベル付けによく使用される「隅付括弧（【 】)」も設定した。

また，テキストマイニングの形態素解析で抽出された単語には複数の類義語が存在する。PLSA で抽出するトピックがよりクリアな内容となるように，目視で類義語辞書を作成し，テキストマイニングのツールに反映した。作成した類義語辞書は，7,240 語の類義語に対して 2,204 語の代表語を登録した。類義語辞書の例を表 1 に示す。表 1 では，背景が灰色の単語が代表語となっている。

表 1　風・空気に関連する特許の要約文から作成した類義語辞書の例

温度調節
温度調節
温度調整
温調
調温
温度設定
中心
中心
中心部
中心側
中心位置
中心部分
筒状
筒状
筒形状
筒体
筒形
筒状体
排気風路
排気風路
排気通路
排気路
排気流路
排気経路

給気風路
給気風路
給気路
給気通路
給気経路
切り替える
切り替える
切り換える
切替える
切換える
方法
方法
手段
本方法
方式
屋外
屋外
室外
室外側
屋外側
吸込空気
吸込空気
吸込み空気
吸い込み空気

螺旋状
螺旋状
渦巻き状
渦巻状
発生
発生
起こる
起きる
使用
使用
利用
使う
熱伝導
熱伝導
熱伝達
熱移動
排ガス
排ガス
排気ガス
排出ガス
噴出
噴出
噴き出す
噴出す

隣接
隣接
隣り合う
隣合う
除霜運転
除霜運転
デフロスト運転
デフロスタモード
行う
行う
行なう
おこなう
濾過
濾過
ろ過
マイナスイオン
マイナスイオン
負イオン
攪拌
攪拌
撹拌
測定結果
測定結果
計測結果

(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。先述したとおり，一般的なPLSAでは，「文書×単語」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×係り受け」という構成をとり，それぞれの単語と係り受けが同時に出現する共起頻度（同時に出現する文章数）を集計したデータを用いる。その構成に採用する単語と係り受けは，特許単位でカウントした頻度が10件以上のものを対象とした。「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。それぞれの共起行列の例を表2, 3に示す。表2, 3は文章単位にカウントした頻度が上位10個の単語と係り受けに限定した共起行列を掲載している。なお，共起行列を作成するにあたり，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（「提供」や「備える」など）や，重要な意味を持たない単語（「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，ストップワードとして共起行列を構成する単語の対象から除外した。ただし，「提供」や「備える」といった単語は係り受けを構成する単語としては含めることとした。

表 2　課題の文章から作成した共起行列の例

	全体の頻度	1,578	1,331	578	548	540	335	296	279	272	264
全体の頻度	単語＼係り受け	空気調和機⇒提供	効率⇒良い	車両用空調装置⇒提供	掃除機⇒提供	容易⇒構成	画像形成装置⇒提供	抑制⇒提供	向上⇒図る	方法⇒提供	装置⇒提供
3,163	空気調和機	1,578	100	4	1	55	0	39	27	2	1
2,802	容易	190	105	51	67	540	28	14	14	32	25
2,700	抑制	142	95	64	63	36	55	296	27	15	7
2,525	良い	113	1,331	12	56	43	31	27	9	19	24
2,339	向上	122	33	24	51	22	8	15	279	11	4
2,062	防止	83	51	23	21	35	34	14	30	6	18
2,042	発生	79	108	40	13	30	43	43	12	13	22
1,745	使用	34	83	14	19	25	3	11	6	23	28
1,717	低減	80	49	44	18	19	10	14	27	16	8
1,582	効率	67	1,331	8	31	26	18	21	7	15	14

表 3　解決手段の文章から作成した共起行列の例

	全体の頻度	1,146	917	776	709	701	695	675	660	625	535
全体の頻度	単語＼係り受け	空気⇒吸い込む	吸い込む⇒空気	連⇒通す	空気⇒吹き出す	備える⇒構成	吹出口⇒吹き出す	空気⇒供給	空気⇒送風	空気⇒排出	制御部⇒備える
7,418	配置	210	191	150	123	145	143	100	127	100	59
4,119	送風機	225	197	112	113	95	73	125	109	58	72
4,060	供給	127	101	67	54	108	53	675	73	81	65
3,830	内部	134	107	126	76	64	44	84	71	111	39
3,738	制御	81	76	29	72	64	103	68	92	47	302
3,660	吹出口	267	301	107	370	108	695	42	63	54	60
3,458	位置	82	72	70	68	72	72	39	44	51	28
3,291	吸い込む	1,146	917	99	313	75	244	68	70	112	58
3,170	発生	118	102	63	80	67	64	49	42	40	47
3,138	外部	142	99	93	70	74	39	62	65	169	36

(3) PLSA の実行と評価

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基準 AIC（Akaike's Information Criterion，赤池情報量基準）[19] で評価して最も評価の良い解を採用することとした。「課題」と「解決手段」のそれぞれの共起行列に対する PLSA の実行解の AIC の評価の結果を図 26, 27 に示す。図 26 の「課題」では，トピック数を 15 から 35 まで 1 刻みで変化させて実行した結果であり，図 27 の「解決手段」では，トピック数を 35 から 60 まで 1 刻みで変化させて実行した結果となっている。各トピッ

ク数に対して 5 つの黒いプロットがあるが，この一つひとつが毎回初期値を変えて PLSA を実行した結果であり，灰色の折れ線グラフはその 5 つの実行解における AIC の平均値を示している。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 [20] という PLSA と同様の分析機能を使用した。

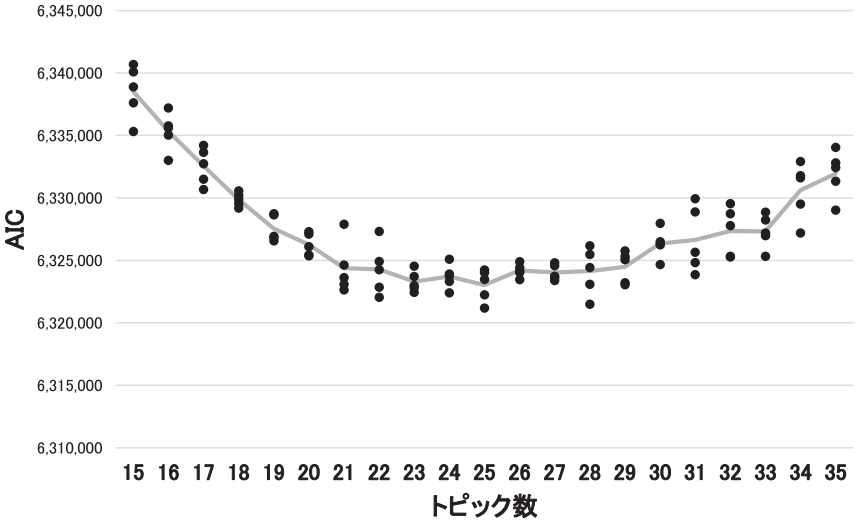


図 26　用途トピック選定のための各トピック数における PLSA の実行解の AIC 評価

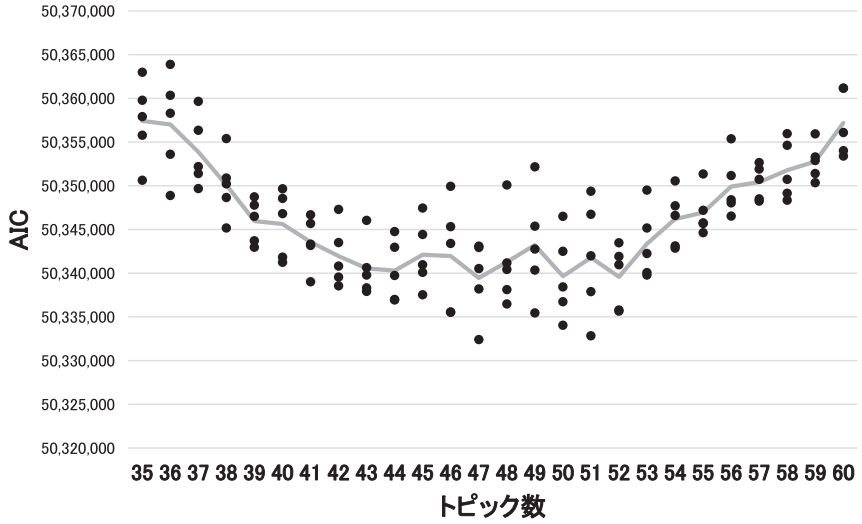


図 27　技術トピック選定のための各トピック数における PLSA の実行解の AIC 評価

図 26, 27 の縦軸の AIC は値が小さいほど良いモデルと評価する指標であるが，その計算式は以下の式（10）で示される。式（10）の第一項は元のデータに対するモデルの当てはまりの良さを示し，第二項はモデルのパラメータの自由度であり複雑度を示す。つまり，モデルの良さを当てはまりの良さだけで評価するのではなく，モデルの複雑度をペナルティとしている。これは，モデルは複雑にするほど（パラメータ数を多くすればするほど），元のデータへの当てはまりは良くなるが，当てはまり過ぎるモデルは一般性を損なう可能性があり，それを第二項のペナルティで抑制しているということである。AIC で評価することで，程良くシンプルで元のデータへの当てはまりも良いモデルを選択することができる。図 26, 27 から AIC の評価はトピック数に対して下に凸のカーブを描いていることがわかる。つまりトピック数を多くするとモデルの当てはまりが良くなるため AIC が改善していくが，あるときからトピッ

ク数の多さにより複雑度のペナルティが効き始め，AIC が悪化していくという現象である。なお，「行 X ×列 Y」の共起行列に PLSA を実行してトピック Z を抽出するモデルの場合，求めるパラメータは P（X|Z），P（Y|Z），P（Z）の 3 つの確率変数であるが，行数を Nx，列数を Ny，トピック数を k とすると，各パラメータの自由度はそれぞれ k（Nx-1），k（Ny-1），k-1 となる。これらの総和が AIC における第二項の自由度となる。

$$AIC = -2(\text{モデルの最大対数尤度}) + 2(\text{モデルの自由度}) \quad (10)$$

採用するトピックの選定の仕方は，まず AIC の平均が最小となるトピック数を定め，その中で個別の AIC が最小となる実行結果を採用した。図 26，27 より，灰色の折れ線が示す各トピック数における AIC の平均値を比較したとき，それが最小となるトピック数は，「課題」から得られる用途トピックについては 25 個，「解決手段」から得られる技術トピックについては 47 個となった。そしてそれぞれのトピック数における 5 つの実行解の中で AIC が最小となる結果を採用した。

（4）トピックの解釈

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各トピックの周辺確率，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 4 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 4（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⇒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。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 4（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⇒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。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 5，表 6 に示す。

表 4　用途トピックと技術トピックの例

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
2.1%	発生	1.4%	加湿機能付空気清浄機⇒提供
2.0%	繁殖	1.3%	ミスト⇒噴霧
1.9%	ミスト	1.3%	繁殖⇒抑制
1.7%	加湿性能	1.2%	十分⇒量
1.5%	ミスト発生装置	1.2%	カビ⇒発生
1.4%	細菌	1.2%	効率⇒良い
1.3%	室内	1.2%	空気調和機⇒提供
1.3%	抑制+できる	1.1%	加湿⇒加湿装置
1.1%	浴室	1.1%	空気⇒加湿
…	…	…	…

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
1.9%	配置	1.0%	送風機⇒吸い込む
1.9%	集塵容器	1.0%	発生⇒送風機
1.6%	旋回	0.9%	含塵空気⇒分離
1.5%	含塵空気	0.9%	備える⇒掃除機
1.4%	捕集	0.8%	掃除機⇒設ける
1.3%	集塵室	0.8%	空気⇒吸い込む
1.3%	通過	0.8%	送風機⇒備える
1.3%	発生	0.8%	掃除機⇒備える
1.2%	集塵装置	0.7%	吸い込む⇒空気
…	…	…	…

表 5　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 6　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		

5.4　トピックのスコアリング

続いて分析対象とした 30,039 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「。」で区切られたセンテンス単位）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（11）と式（12）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（11）のP(Sx|X）はN(X）の逆数,式（12）のP(Sy|Y）はN(Y）の逆数として計算される。式（11）のP(X|T)と式（12）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（13）で計算されるが，SxとSyは文Sにおいて重みは同じであるため，P(S|Sx）とP(S|Sy)それぞれ0.5とする。またP(S）は式（14）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周辺確率に該当する。

$$\mathrm{P}(Sx \mid T) = \sum\nolimits_{X} \mathrm{P}(Sx \mid X)\mathrm{P}(X \mid T) \tag{11}$$

$$\mathrm{P}(Sy \mid T) = \sum\nolimits_{Y} \mathrm{P}(Sy \mid Y)\mathrm{P}(Y \mid T) \tag{12}$$

$$\mathrm{P}(S \mid T) = \mathrm{P}(S \mid Sx)\mathrm{P}(Sx \mid T) + \mathrm{P}(S \mid Sy)\mathrm{P}(Sy \mid T) \tag{13}$$

$$\mathrm{P}(S) = \sum\nolimits_{T} \mathrm{P}(S \mid T)\mathrm{P}(T) \tag{14}$$

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7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。つまりこれはBag-of-Topicsという形式のデータであり，PLSAでトピックを抽出しそのスコアを計算することで，特許文書の記述情報が集約された新たな分析軸が追加された。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することができる。なお，この先の各分析はすべてこのデータセットをベースとしている。

表7　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

5.5　トピックのトレンド分析

表7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。この指標はトピックのスコアリングで用いた事後確率と事前確率の比率を示すリフト値となってい

る。このリフト値はトピックTを条件とすることで出願年Yrの元々の出願件数の割合が何倍になるのかを示している。各トピックにおいて，ある出願年の出願件数の割合が全体の割合よりも大きいほど値は高くなる。つまりこのリフト値は，各年における元々の出願件数の大小の影響を除した形で，トピックと出願年の関係の強さを表している。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年でこのリフト値の上昇率が高い上位5つのトピックのトレンドをそれぞれ図28, 29に示す。

図28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29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。

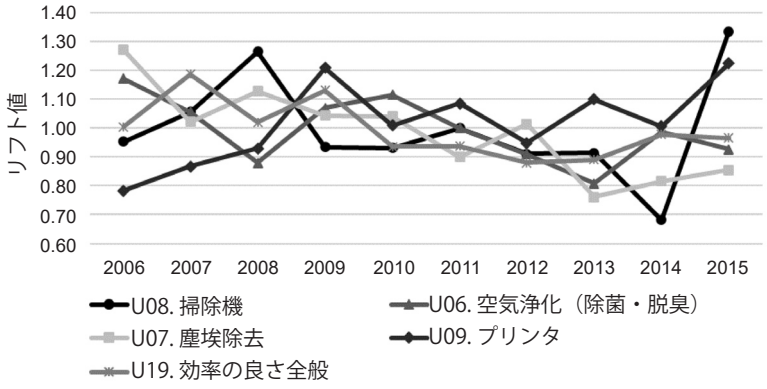


図28　2013年からの上昇率トップ5の用途トピックのトレンド

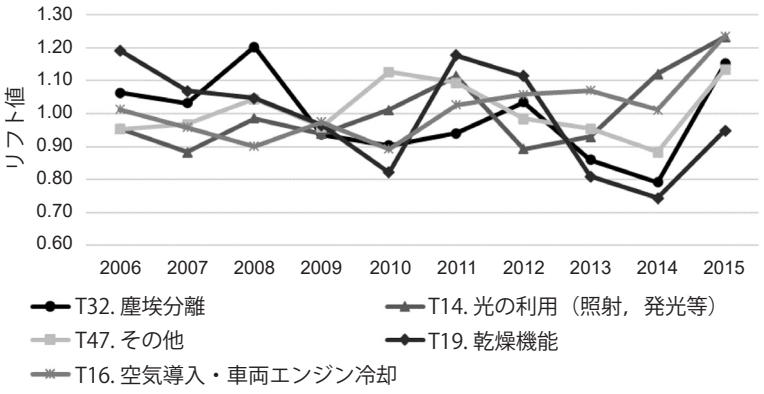


図29　2013年からの上昇率トップ5の技術トピックのトレンド

5.6　トピックを用いた競合他社の分析

表7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。

(1)　静的分析：出願人のポジショニングマップ

本分析事例では，まず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，独自の特有な技術を保有している可能性がある。

ここでは先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30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30より，まずシェアの高さを確認するとA社，B社，C社が高く，特にC社は注力度が非常に高いため，C社は特有の技術力があると考えられる。一方，E社，G社，I社はシェアは中程度だが，注力度は比較的高いため，技術力もあると考えられる。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，より技術力を高めながらシェアを伸ばすことが期待できる。あるいは中程度のシェアの企業の間で連携をすることで，業界大手に対抗するということも考えられる。

このように塵埃分離に関する技術は，3社のシェアが高いものの，他にもある程度のシェア・注力度を持つ企業が何社か存在し，また先のトレンド分析でも近年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30　技術トピック「T32. 塵埃分離」における出願人のポジショニングマップ

(2) 動的分析：出願人の出願件数の推移

こうした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る。つまり，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「T32. 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31に示す。

図31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，たとえば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社である。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。こうした結果からさまざまな技術戦略を検討するヒントが得られると期待できる。

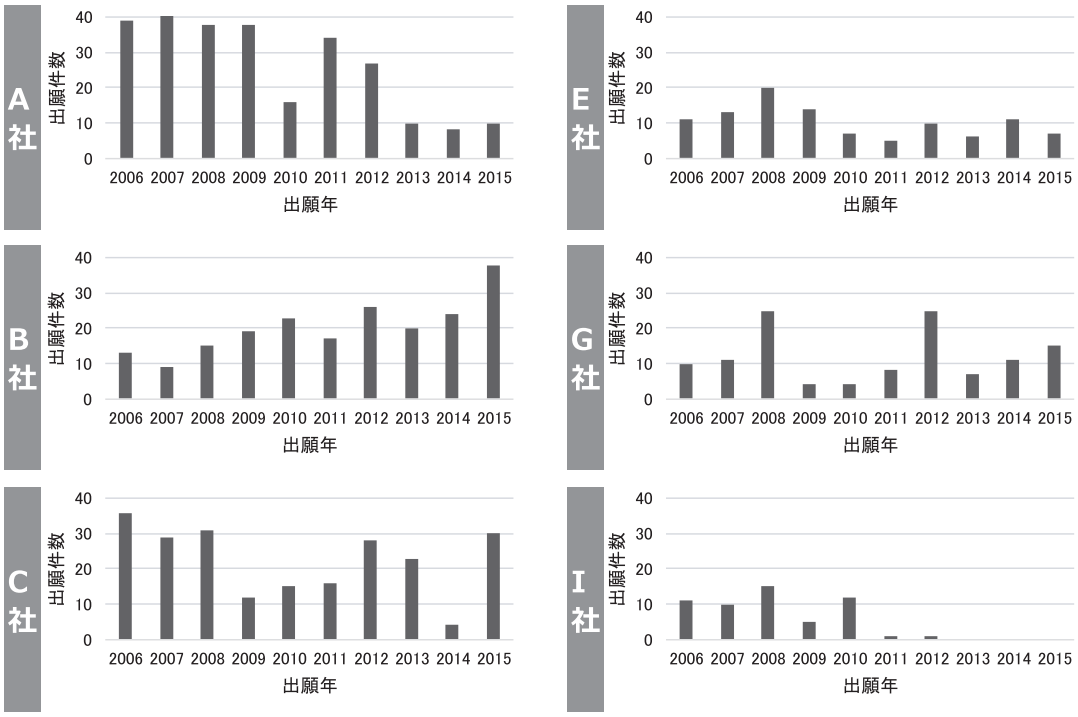


図 31　技術トピック「T32. 塵埃分離」における出願人の出願件数の推移

5.7　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 7 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，それぞれは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。

5.7.1　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 32 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）を使用した。

（1）各用途と関係する技術の把握

図 32 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シミュレーションにより，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できる。ここでは先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 32 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 33 に示す。図 33 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）を示している。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わかる。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。

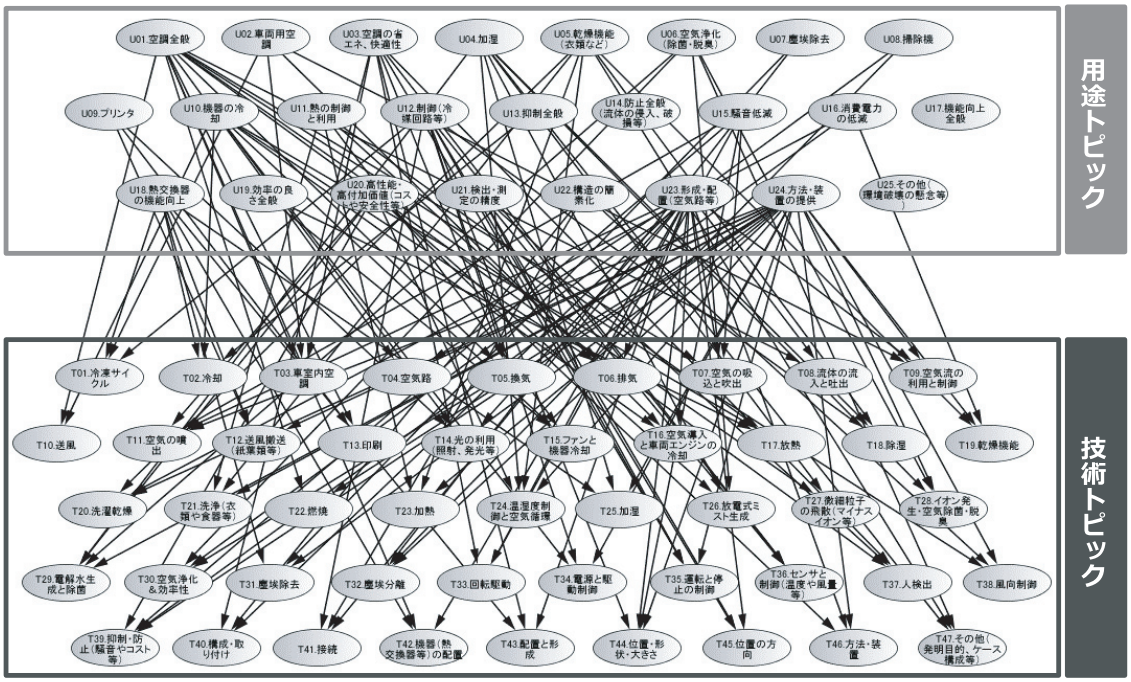


図 32　ベイジアンネットワークを適用した用途⇒技術の関係モデル

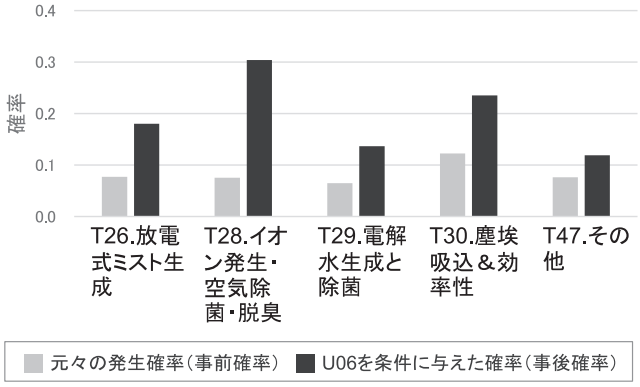


図 33　用途トピック「U06. 空気浄化」を条件に与えたときに確率が上昇する技術トピック

（2）用途×技術における競合他社の動向の把握

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，先ほどと同様の競合他社の分析を実施した。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 34 に示す。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わかる。「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，「U06. 空気浄化（除菌・脱臭）」の用途を実現しようとしたとき，たとえば一強状態の技術を避けるのであれば，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 で買収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，2013 年に業務用として，2017 年には家庭用として発売されている。つまり，その家電のコア技術となる電解水を生成する技術は 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。

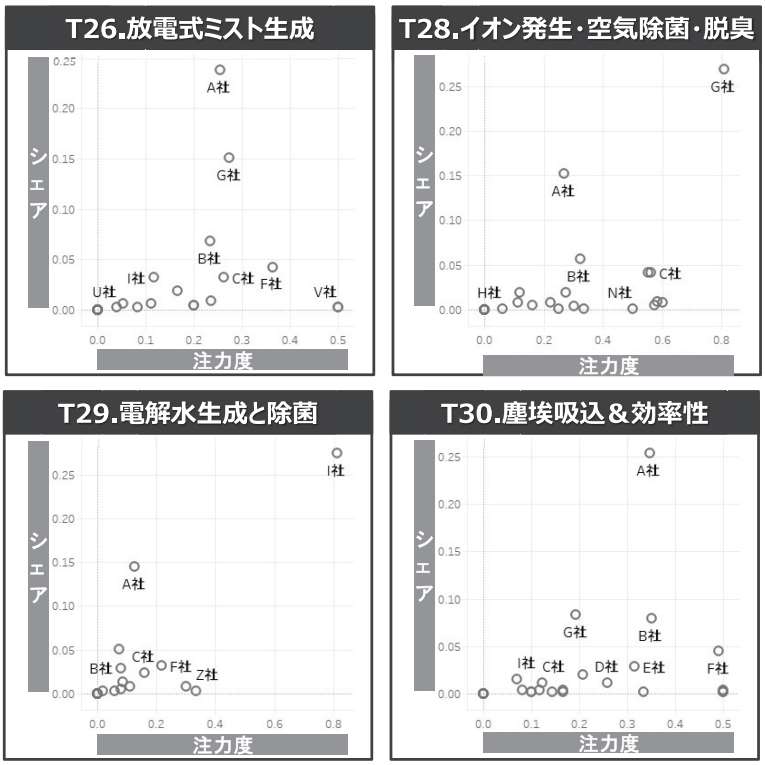


図 34　用途トピック「U06. 空気浄化」と関係のある技術トピックの出願人ポジショニングマップ

5.7.2　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 35 に構築されたモデルの結果を示す。

（1）各技術と関係する用途の把握

図 35 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。ここでは技術トピック「T18. 除湿」と「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 35 のモデルの結果より，技術トピック「T18. 除湿」と「T32. 塵埃分離」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それぞれ図 36, 37 に示す。図 36, 37 のグラフでは，技術トピック T18 あるいは T32 と関係が見られた各用途トピックの元々の確率（事前確率）と，技術トピック T18 あるいは T32 を条件に与えたときの条件付確率（事後確率）を示している。どの用途トピックも事後確率の方が高くなっているため，これらの用途と技術の関係が強いことがわかる。特に「T32. 塵埃分離」に対して関係が見られた用途トピックはただ一つ「U08. 掃除機」のみであり，事後確率の上昇の仕方も顕著である。

さらにここから，技術「T18. 除湿」×用途「U05. 乾燥機能（衣類など）」の関係と，技術「T32. 塵埃分離」×用途「U08. 掃除機」の関係を取り上げ，技術の新しい用途展開を探索する考え方について紹介する。図 36, 37 の結果から，当然 T18 の除湿の技術は U05 の衣類などの乾燥機能の用途を想定して出願されている特許が多く，T32 の塵埃分離の技術は U08 の掃除機の用途を想定して出願されている特許が多いということになる。しかし，この傾向に当てはまらない特許も存在している。技術の新しい用途展開を探索する上ではそうした特許がヒントになり得る。以下に具体的な探索例を紹介する。

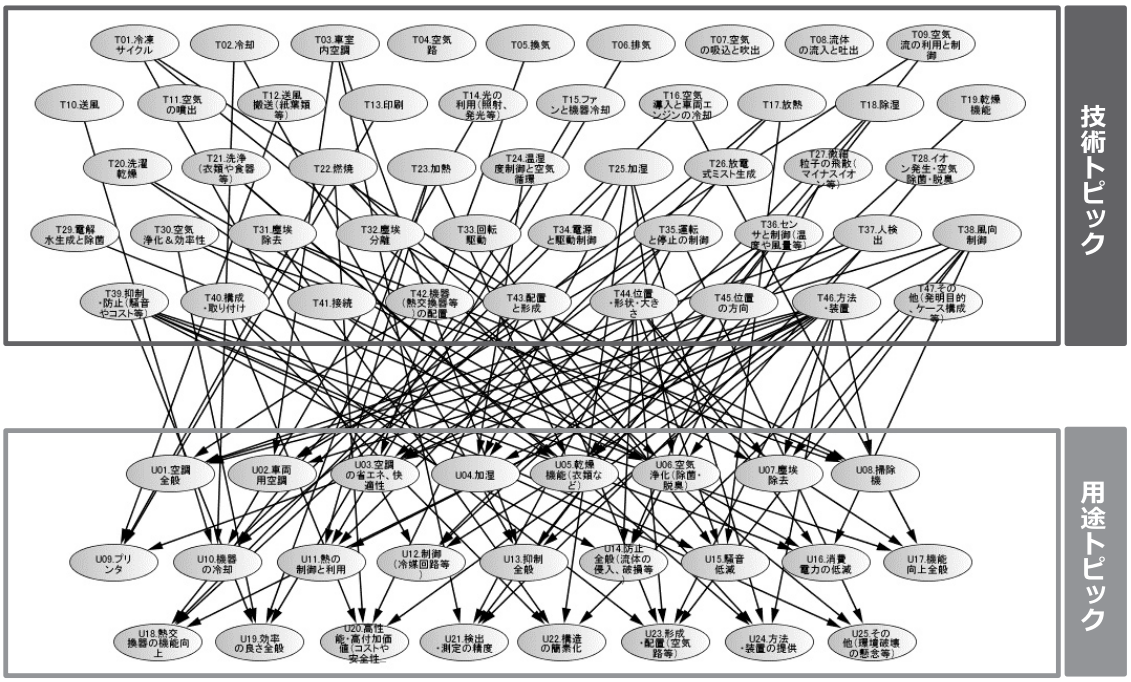


図 35　ベイジアンネットワークを適用した技術⇒用途の関係モデル

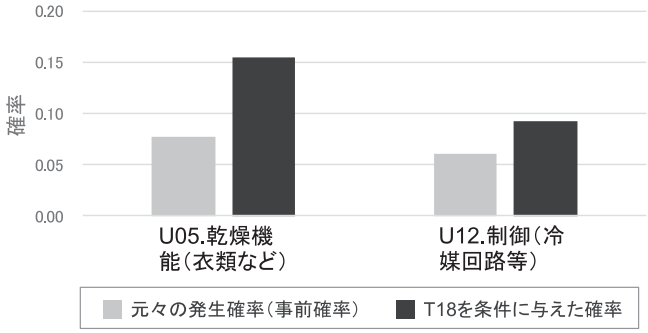


図 36　技術トピック「T18. 除湿」を条件に与えたときに確率が上昇する用途トピック

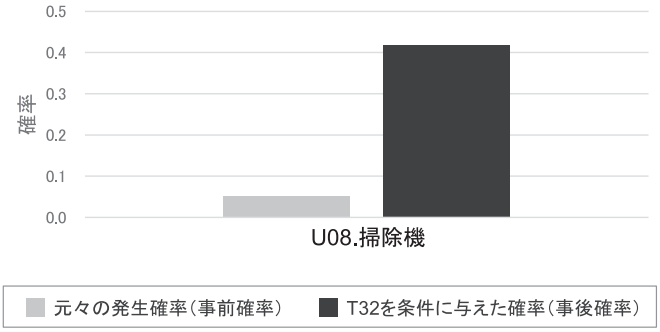


図 37　技術トピック「T32. 塵埃分離」を条件に与えたときに確率が上昇する用途トピック

（2）技術「T18. 除湿」×用途「U05. 乾燥機能（衣類など）」から探索する新規用途

まず「T18. 除湿」の技術が「U05. 乾燥機能（衣類など）」の用途を想定して出願されている特許の要約文の典型例を図 38（左）に示す。これはドラム式洗濯乾燥機の特許である。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る。一方，「T18. 除湿」の技術が「U05. 乾燥機能（衣類など）」の用途を想定しないで出願されている特許の要約文の例を図 38（右）に示す。これはインクジェットプリンタに関する特許である。特許の要約の内容を確認すると，インク液を吸収した紙の湿気

T18の技術がU05の用途を想定して出願された特許例	T18の技術がU05の用途を想定せず出願された特許例
【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。プレ加熱部は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，均一な膜厚を持つ固体状の凝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る

図 38　技術「T18. 除湿」が応用されている 2 つの用途の特許例

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る。この 2 つの特許の比較から，たとえば，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，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ることができる。

（3）技術「T32. 塵埃分離」×用途「U08. 掃除機」から探索する新規用途

同様に技術「T32. 塵埃分離」と用途「U08. 掃除機」の関係についても探索していく。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要約文の典型例を図 39（左）に示す。これはサイクロン掃除機の特許である。特許の要約の内容を確認すると，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る。一方，「T32. 塵埃分離」の技術が「U08. 掃除機」の用途を想定しないで出願されている特許の要約文の例を図 39（右）に示す。これは画像形成装置，つまりプリンタに関する特許である。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処理を実行することで，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。この 2 つの特許の比較から，たとえば，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，センサで判断して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。

T32の技術がU08の用途を想定して出願された特許例	T32の技術がU08の用途を想定せず出願された特許例
【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る

図 39　技術「T32. 塵埃分離」が応用されている 2 つの用途の特許例

（4）イノベーションの鍵となるアイデアの創出

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。

5.8　Nomolytics を適用した特許文書データ分析のまとめ

分析事例の最後として，ここでは Nomolytics を適用した特許文書データ分析のメリットを二つにまとめる。

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ある。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，それを解釈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ができる。これにより，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できる。

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ある。従来の特許分析でも，図 22 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ったが，「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，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る。

このように，従来の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 技術を組み合わせた Nomolytics で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。今回は特許の要約文を分析対象としており，最終的には特許文書の明細まで詳細に確認することが求められるが，こうした要約文のトピックをベースにした分析を進めることで，詳細まで確認すべき重要な特許の対象を効率的に絞り込んでいくことができる。

6.　Nomolytics を特許文書データに適用した分析事例（その 2）

Nomolytics を実際の特許文書データに適用した分析事例もう一つ紹介する。なお，二つ目の事例も特許の要約文を分析対象としているが，一つ目の事例のように要約文を「課題」の文章と「解決手段」の文章に分けてトピックを抽出するのではなく，今回は要約文の全文を対象にトピックを抽出する。また，一つ目の事例で適用したベイジアンネットワークはここでは適用せず，Nomolytics のアプローチの中でも PLSA でトピックを抽出し，そのトピックを軸に傾向を分析するまでの事例を紹介する。

6.1　分析で用いるデータ

本分析事例では，特許の要約と請求項に「車」「電気」という 2 つのキーワードを含む国内の特許公報データ 26,419 件を分析対象としている。出願期間は 2007 年 1 月 1 日から 2016 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。今後開発と普及の拡大が見込まれる「電気自動車」に関連した技術を分析する狙いから抽出したデータである。

6.2　分析の全体像

今回は特許の要約文全文を対象にトピックを抽出し，その傾向を分析するが，分析プロセスの全体像は一つ目の分

析事例の図 25 で示した（4）のステップまでが該当する。すなわち（1）テキストマイニング× PLSA によるトピックの抽出，（2）トピックのスコアリング，（3）トピックのトレンドの分析，（4）トピックを用いた競合他社の分析というステップである。要約文を「課題」と「解決手段」に分割してトピックを抽出すること以外は，基本的に一つ目の分析事例と同様のアプローチで分析を進める。以下に分析ステップごとに説明する。

6.3　テキストマイニング× PLSA による要約トピックの抽出

テキストマイニングと PLSA を用いて特許の要約文全体の内容をいくつかのトピックに集約する。そのアプローチは基本的に一つ目の分析事例と同様であるが，以下にトピック抽出の手順を述べる。

（1）テキストマイニング

要約文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。単語は今回は名詞のみを抽出し，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およびその逆の係り受け（動詞・形容詞・形容動詞に対する名詞）の単語ペアを抽出する。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）を使用した。本ツールの形態素解析において，文章の区切りを定義する文字列には「句点（。）」の他に，特許要約文の中で文章のラベル付けによく使用される「隅付括弧（【 】）」も設定した。

一つ目の分析事例と同様に，テキストマイニングの形態素解析で抽出された単語に対して，目視で類義語辞書を作成し，テキストマイニングのツールに反映した。作成した類義語辞書は，3,278 語の類義語に対して 1,343 語の代表語を登録した。類義語辞書の例を表 8 に示す。表 8 において，背景が灰色の単語が代表語となっている。

表 8　電気自動車に関連する特許の要約文から作成した類義語辞書の例

充電スタンド
充電スタンド
充電ステーション
充電設備
充電施設
給電設備
給電スタンド
給電装置
給電装置
送電装置
電力供給装置
電源供給装置
給電機器
省エネ
省エネ
省エネルギー
省エネルギー化
省電力
省電力化
サスペンション
サスペンション
サスペンション装置
懸架
懸架装置

簡素化
簡素化
簡略化
単純化
簡易化
先端
先端
先端部
先端側
先端部分
被覆
被覆
コーティング
被膜
皮膜
電子機器
電子機器
電子装置
電子機器用
電子デバイス
固定子巻線
固定子巻線
ステータコイル
ステータ巻線

フランジ
フランジ
フランジ部
鍔部
誤検出
誤検出
誤検知
誤判定
小型化
小型化
小型軽量化
コンパクト化
送風機
送風機
送風装置
ブロワ
漏洩
漏洩
漏れ
漏出
熱伝導
熱伝導
熱伝達
伝熱

Li
Li
リチウム
SMR
SMR
システムメインリレー
車軸
車軸
アクスル
共重合体
共重合体
コポリマー
羽根車
羽根車
インペラ
接地線
接地線
アース線
緊急時
緊急時
非常時
僅か
僅か
わずか

（2）共起行列の作成

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。共起行列は「単語×係り受け」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（同時に出現する文章数）を集計したデータを用いる。なお共起行列の構成に採用する単語と係り受けは，特許単位でカウントした頻度が 20 件以上のものを対象とし，「名詞単語（3,020 語）×係り受け（2,128 表現）」の共起行列を作成した。共起行列の例を表 9 に示す。表 9 は文章単位にカウントした頻度が上位 10 個の単語と係り受けに限定した共起行列を掲載している。なお，共起行列を作成するにあたり，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（「課題」「解決手段」「選択図」「提供」など）や，重要な意味を持たない単語（「前記」「本発明」「当該」「個」など）は，トピック抽出においてノイズになり得るため，ストップワードとして対象から除外した。

表 9　電気自動車関連の要約文から作成した共起行列の例

全体の頻度	係り受け / 単語	電力⇒供給	否⇒判定	バッテリ⇒充電	モータ⇒駆動	効率⇒良い	供給⇒電力	充電⇒行う	電気自動車⇒提供	並列⇒接続	モータ⇒供給
全体の頻度		1,350	539	529	494	444	368	306	289	282	280
5,880	構成	118	33	33	36	24	32	25	10	46	30
5,188	モータ	239	61	58	494	31	54	19	33	27	280
5,092	制御	268	73	85	108	12	115	41	2	40	74
4,655	電気自動車	193	73	129	79	59	53	114	289	13	54
4,604	配置	69	2	8	29	6	15	9	5	12	18
4,602	バッテリ	337	87	529	60	44	81	72	20	70	106
4,006	形成	31	4	8	20	1	7	5	0	12	4
3,978	供給	1,350	43	85	53	10	368	43	7	43	280
3,960	検出	134	99	36	56	4	44	19	8	27	33
3,629	電力	1,350	50	127	75	35	368	57	20	36	189

（3）PLSA の実行と評価

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。PLSA の実行は一つ目の分析事例と同様にトピック数を 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基準 AIC で評価して最も評価の良い解を採用する。PLSA の実行解の AIC の評価の結果を図 40 に示す。図 40 では，トピック数を 20 から 45 まで 1 刻みで変化させて実行した結果となっている。各トピック数に対して 5 つの黒いプロット一つひとつが初期値を変えて PLSA を実行した結果であり，灰色の折れ線グラフはその 5 つの実行解における AIC の平均値を示している。灰色の折れ線が示す AIC の平均値を比較したとき，それが最小となるトピック数は 34 個となり，このトピック数 34 個における 5 つの実行解の中で AIC が最小となる結果を採用した。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を使用した。

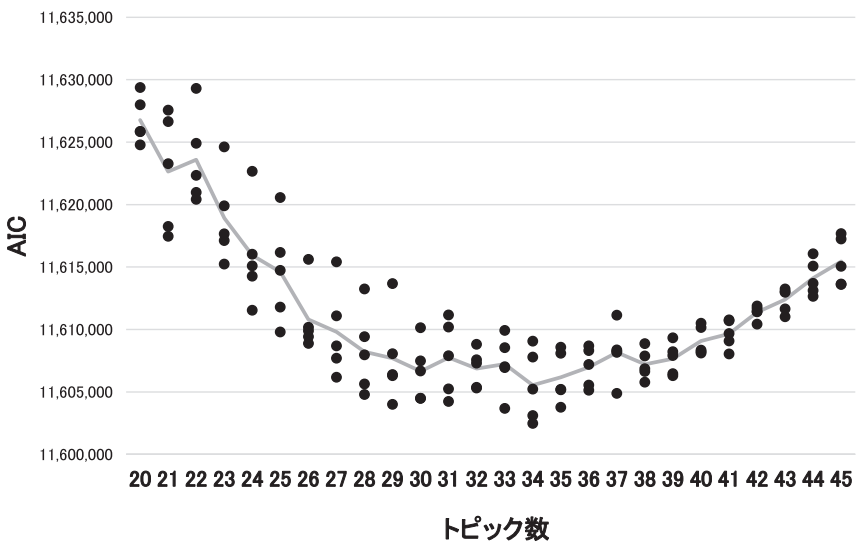


図 40　要約トピック選定のための各トピック数における PLSA の実行解の AIC 評価

（4）トピックの解釈

抽出されたトピックの内容の例を表 10 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 10（左）のトピック Z02 では，単語は，連結，伝達，モータ，出力軸，トランスミッション，クラッチなどで所属確率が高く，係り受けは，介する⇒伝達，介する⇒連結，動力⇒伝達，出力軸⇒連結，駆動力⇒伝達，トルク⇒伝達といった表現で所属確率が高い。つまり，この結果は動力の伝達に関するトピックであると解釈できる。表 10（右）のトピック T13 では，単語は，充電，電気自動車，蓄電装置，バッテリ，充電システム，蓄電池などで所属確率が高く，係り受けは，バッテリ⇒充電，充電⇒行う，電気自動車⇒充電，蓄電池⇒充電，蓄電装置⇒充電，電力⇒供給といった表現で所属確率が高い。つまり，この結果は電気自動車の蓄電池充電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全 34 個の要約トピックの一覧を表 11 に示す。

表 10　電気自動車関連の要約トピックの例

トピックZ02			
確率	単語	確率	係り受け
6.8%	連結	2.2%	介する⇒伝達
4.7%	伝達	2.0%	介する⇒連結
3.6%	モータ	1.9%	動力⇒伝達
3.0%	出力軸	1.8%	出力軸⇒連結
2.8%	トランスミッション	1.8%	駆動力⇒伝達
2.6%	クラッチ	1.4%	トルク⇒伝達
2.1%	内燃機関	1.3%	クラッチ⇒介する
2.0%	駆動部	1.2%	駆動輪⇒伝達
2.0%	入力軸	1.1%	クラッチ⇒備える
1.7%	駆動輪	1.1%	出力軸⇒接続
1.6%	駆動軸	1.1%	機械的⇒連結
1.6%	動力	1.0%	モータ⇒備える
1.6%	駆動	1.0%	トランスミッション⇒備える
1.4%	トルク	0.9%	モータ⇒連結
1.3%	駆動装置	0.9%	駆動軸⇒連結
…	…	…	…

トピックZ13			
確率	単語	確率	係り受け
12.6%	充電	5.1%	バッテリ⇒充電
8.9%	電気自動車	4.0%	充電⇒行う
6.5%	蓄電装置	3.9%	電気自動車⇒充電
3.0%	バッテリ	1.9%	蓄電池⇒充電
2.0%	充電システム	1.6%	蓄電装置⇒充電
2.0%	蓄電池	1.6%	電力⇒供給
1.9%	電力	1.3%	充電⇒開始
1.7%	制御	1.2%	電気自動車⇒接続
1.5%	充電スタンド	1.2%	充電⇒蓄電装置
1.5%	放電	1.1%	用いる⇒充電
1.3%	外部電源	1.1%	充電⇒制御
1.2%	充電+できる	1.0%	供給⇒電力
1.0%	充電ケーブル	0.9%	蓄電装置⇒備える
0.8%	検出	0.9%	電力⇒充電
0.7%	情報	0.8%	蓄電装置⇒提供
…	…	…	…

表 11　電気自動車関連の要約トピックの一覧

No.	トピック名	No.	トピック名
Z01	エンジンの始動と停止	Z18	演算・推定
Z02	動力の伝達	Z19	機器の異常検出
Z03	モータ駆動	Z20	操作スイッチ
Z04	ロータ・ステータなど回転部品の構成	Z21	筐体
Z05	ブレーキ装置	Z22	表面の形成
Z06	動作制御	Z23	位置とその移動
Z07	動力伝達の制御	Z24	配置・位置・方向
Z08	スイッチの切り替え	Z25	構成の方位
Z09	交流・直流の変換	Z26	構成
Z10	エネルギーの変換	Z27	接続
Z11	電池モジュールの提供	Z28	方法の提供
Z12	二次電池の構成	Z29	損傷や浸水など不具合の防止
Z13	電気自動車の蓄電池充電	Z30	小型化・簡素化・低コスト化など付加価値
Z14	非接触受電など給電装置	Z31	効率性・安全性の向上
Z15	外部への電力供給	Z32	既存エンジンへの警鐘・樹脂組成物の提供
Z16	空調などの冷却・加熱	Z33	重力発電の活用による地球温暖化防止
Z17	情報通信	Z34	タービン発電の出力向上・燃費低減

6.4　トピックのスコアリング

続いて分析対象とした 26,419 件の特許データに対して，今回抽出された 34 個の要約トピックのスコア（該当度）を計算した。スコアの計算方法は一つ目の分析事例と同様であるため割愛する。連続値であるスコアから該当有無を示す 1,0 のフラグに変換する閾値は，スコアの分布や実際の文章の内容も確認しながら 5 に設定した。トピックのスコアリングの計算処理により，表 12 に示すようなデータが作成された。26,419 件の特許データは，出願年，出

願人，要約文という元々の情報に加え，要約のトピック 34 個の 1,0 のフラグ情報が付加されたデータとなり，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することができる。この先の各分析はすべてこのデータセットをベースとしている。

表 12　要約トピックの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	トピック Z01	トピック Z02	…	トピック Z34
1	特願2007-XXXX	2007	A社	【課題】電気式変速操作装	1	1		0
2	特願2009-XXXX	2009	B社	【課題】従来の電気自動車	0	1		1
3	特願2012-XXXX	2012	C社	エンジンのための方法及び	0	1		1
4	特願2013-XXXX	2013	D社	【課題】駐車場に設置され	1	0		0
…	…	…	…		…	…		…
26,419	特願2016-XXXX	2016	X社	充電ステーションが電気エ	1	0		1

6.5　トピックのトレンド分析

表 12 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。一つ目の分析事例と同様に出願年とトピックの関係の強さを示すリフト値の経年変化を可視化する。たとえば，2014 年からの直近 3 年でリフト値の上昇率が高い上位 4 つのトピックのトレンドを図 41 に示す。

図 41 より，特に「Z08. スイッチの切り替え」が上昇しており，他には「Z01. エンジンの始動と停止」や「Z19. 機器の異常検出」に関するトピックが上昇している。

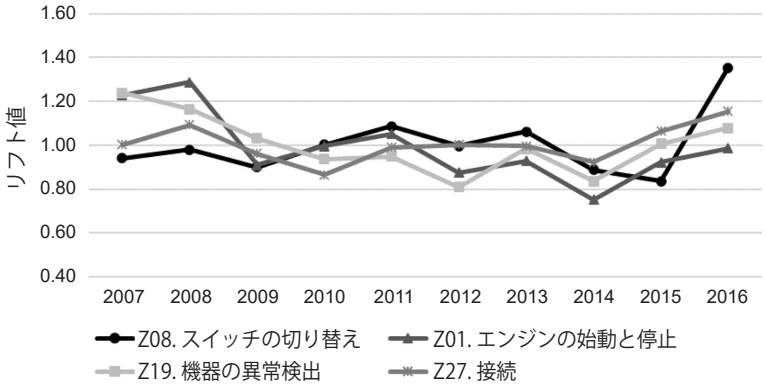


図 41　2014 年からの上昇率トップ 4 のトピックのトレンド

6.6　トピックを用いた競合他社の分析

表 12 のトピックのフラグデータを用いて，各トピックにおける出願人の特徴を分析する。

（1）静的分析：出願人のポジショニングマップ

一つ目の分析事例と同様に，「シェア」と「注力度」という 2 つの指標を縦軸と横軸に設定し，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する。

ここでは先のトレンド分析で最も上昇していたトピック「Z08. スイッチの切り替え」を例とした結果を図 42 に示す。図 42 より，まずシェアの高さを確認すると，A 社が圧倒的に高く，この領域で多くの特許を出願している。一方，注力度を確認すると，O 社が最も高く他社とは大きなギャップがある。O 社はこの領域に全社的に関心が高く，特有の技術力を保有している可能性が考えられる。また，中程度のシェア・注力度には B 社，D 社，H 社などがあるが，こうした企業の間で連携することで，より技術力を高めながらシェアを伸ばし，A 社という巨人に対抗することも考えられる。

（2）動的分析：出願人の出願件数の推移

ここで，注目の対象としたA社，B社，D社，H社，O社について，一つ目の分析事例と同様に，この「Z08. スイッチの切り替え」というトピックに該当する特許の出願件数の推移も可視化した。その結果を図43に示す。図43より，シェア1位のA社は，ここ最近は減少傾向にあったが，直近では出願が増加している。シェア2位のB社は，最近は減少傾向にあり，シェア3位のD社も最近は出願件数が少なくなっている。H社と，注力度1位のO社は，他と比べると件数は全体的に少ないが，どちらも直近で出願が増えている。ここから，直近で出願件数を急に増やしているA社，H社，O社は，技術戦略の転換の可能性も考えられるため，直近での出願特許の内容や今後の出願動向には要注目といえる。

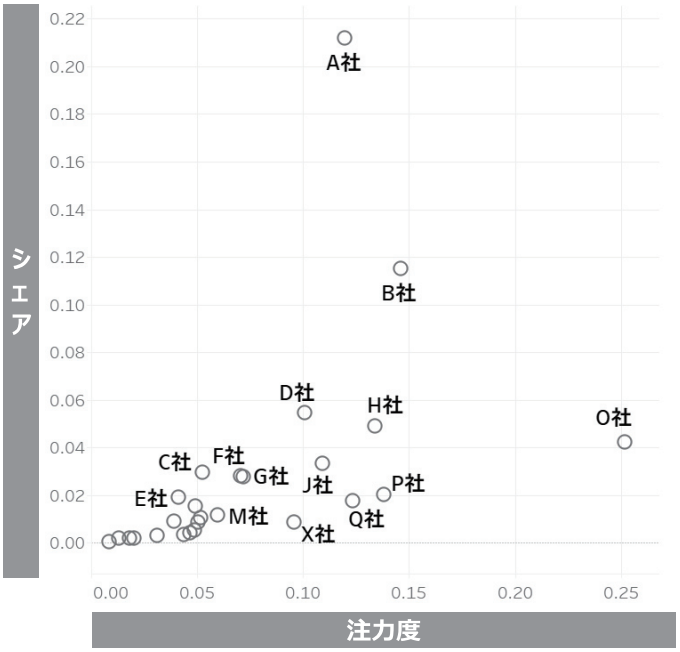


図42　トピック「Z08. スイッチの切り替え」における出願人のポジショニングマップ

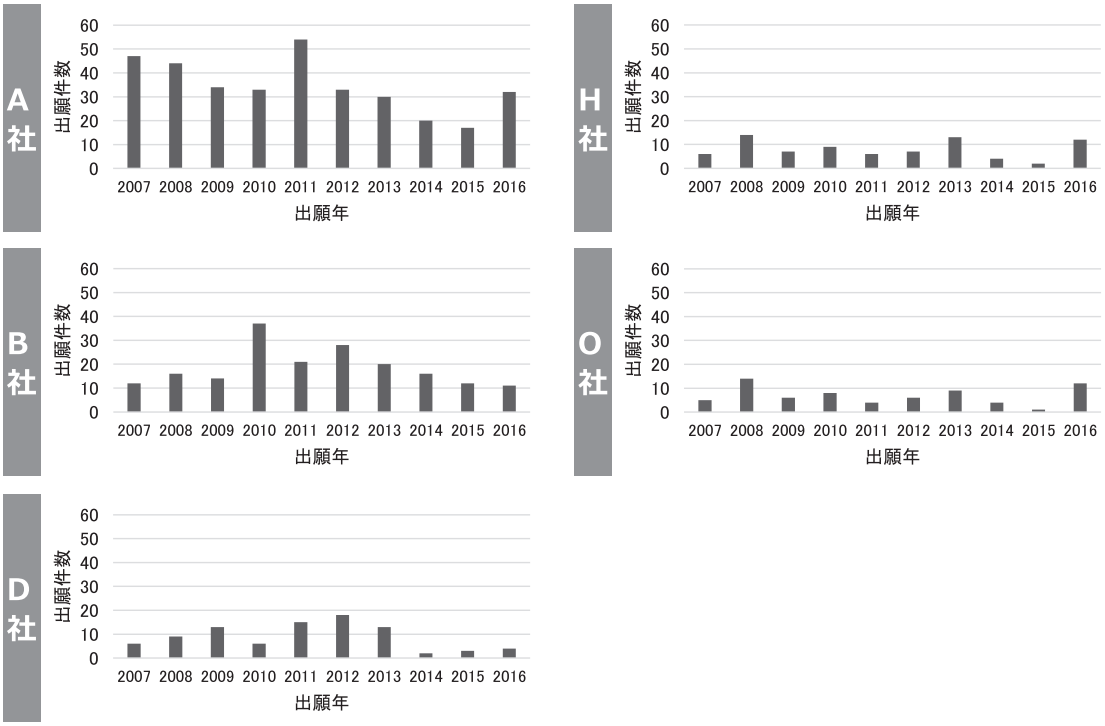


図43　トピック「Z08. スイッチの切り替え」における出願人の出願件数の推移

（3）出願動向を示す特許内容の確認

直近で出願件数が増えている A 社，H 社，O 社の 3 社について，その直近一年（2016 年）に出願している特許の内容を確認した。

まず A 社が 2016 年に Z08 のトピックの特許を出願した件数は 32 件あったが，そのうち 8 件の特許を例に，「発明の名称」と「要約の課題」の部分を抜粋したものを図 44 に示す。今回確認した特許は，どれも「Z08. スイッチの切り替え」というトピックに該当している特許であるが，たとえば，図 44 の左側の特許①～④は，電気自動車の衝突時にコンデンサを放電に切り替える技術に関する特許となる。これは衝突時の感電を防止する重要な技術となる。また，図 44 の右側の特許⑤～⑧は，何か異常が生じたときにそれを検知して回避する切り替え技術に関する特許となる。

続いて，H 社と O 社が 2016 年に出願した Z08 のトピックの特許も確認したが，実は H 社と O 社の 2016 年の出願特許はすべて H 社と O 社の共同出願であった。この共同出願の件数は 12 件あったが，そのうち 6 件の特許を例に，「発明の名称」と「要約の課題」の部分を抜粋したものを図 45 に示す。たとえば，図 45 の左側の特許①～③は，電力の供給が途絶えた場合でも給電を維持できる切り替え技術に関する特許となる。また，図 45 の右側の特許④～⑥は，蓄電部などに異常が発生したときに，他にその影響が及ぶことを回避する切り替え技術に関する特許となる。このように，H 社と O 社の 2 社はすでに連携することでこの切り替え技術の領域において開発に力を入れていることが確認できる。

Nomolytics を特許文書データに適用した二つ目の分析事例は以上となる。より実践的な特許分析では，さらにここから特許の明細まで深く確認していくことが重要となるが，このようにトピックをベースにしたシンプルで分かりやすい分析をドリルダウンで進めていくことで，注目すべき特許を効率的に発見することができる。

出願特許 ①	
発明の名称	電気自動車
要約【課題】	車両の衝突時に平滑化コンデンサを放電する確実性を向上させる。

出願特許 ②	
発明の名称	電気自動車用の電源システム
要約【課題】	車両が衝突したときにより確実に平滑化コンデンサを放電する。

出願特許 ③	
発明の名称	電気自動車
要約【課題】	電気自動車の衝突時に，できるだけパーキングロックを使わずにモータを停止させて平滑化コンデンサを放電させる。

出願特許 ④	
発明の名称	ハイブリッド車両
要約【課題】	車両の衝突時に，モータの回転数を取得できない場合にも，インバータに接続されるコンデンサの電荷の放電を速やかに完了する。

出願特許 ⑤	
発明の名称	電気自動車
要約【課題】	補機バッテリが短絡しても，重要な補機への電力供給を継続できる電気自動車を提供する。

出願特許 ⑥	
発明の名称	電気自動車用の電源システム
要約【課題】	通信不良が生じた場合であってもシステムスイッチを安全に開放することのできる電気自動車用の電源システムを提供する。

出願特許 ⑦	
発明の名称	電気自動車
要約【課題】	第1インバータ回路と第2インバータ回路のうち一方で異常が生じたときに，両インバータ回路の複数のスイッチング素子を同時にオフするモータ制御ユニットを提供する。

出願特許 ⑧	
発明の名称	ハイブリッド車
要約【課題】	ハイブリッド車の第1コントローラに異常が生じたときでも，エンジンを始動し得る技術を提供する。

図 44　A 社が 2016 年に出願した「Z08. スイッチの切り替え」の特許例

出願特許 ①	
発明の名称	給電中継回路，副電池モジュール，電源システム
要約【課題】	一つの車載電源が失陥した場合であっても，車両の電気的負荷に給電する。

出願特許 ②	
発明の名称	車載用のバックアップ装置
要約【課題】	電源部からの電力供給が途絶えた場合であっても電力供給対象への電力供給を途切れさせることなく供給源を蓄電部に切り替えることが可能な装置を，より簡易な構成で実現する。

出願特許 ③	
発明の名称	車両用電源装置
要約【課題】	第１電源部からの電力に基づく急速充電動作と，充電中に第１電源部の電力供給が遮断されても遮断前後で放電状態を維持し得る充放電動作とを行い得る車両用電源装置を提供する。

出願特許 ④	
発明の名称	車両用電源装置
要約【課題】	第１電源部の電力不足によって特定負荷が駆動できなくなる事態を，第２電源部に及ぼす影響を抑えて回避し得る車両用電源装置を提供することを目的とする。

出願特許 ⑤	
発明の名称	リレー装置及び車載システム
要約【課題】	少なくとも２つの蓄電部に対して発電機から充電電流を供給することができ，且つ一方の蓄電部側に異常が発生した場合に，その異常が他方の蓄電部側に及ぶことを抑え得るリレー装置を提供する。

出願特許 ⑥	
発明の名称	車両用電源装置
要約【課題】	アイドリングストップ状態からの復帰の際に発電機側の蓄電部の電圧が低下しても負荷にその影響が及びにくい車両用電源装置を，電流集中を抑制し得る構成で実現する。

図 45　H 社と O 社が 2016 年に共同出願した「Z08. スイッチの切り替え」の特許例

7.　インサイト獲得のための PLSA の展開技術

ここからは，より深いインサイトを獲得できるようなトピック抽出の方法として，PLSA を応用して筆者が開発した“PCSA”と“differential PLSA”という２つの開発技術について紹介する。

これはテキストデータの分析に限った話ではないが，ビジネスにおけるデータ分析では，新たな気づきとなるインサイトを抽出できることがしばしば求められる。しかし，いざデータ分析に取り組むものの，結果が期待外れということは少なくない。それは，その業務担当者からすると経験的によく知っている結果であり，目新しさがないということが大半である。この期待と現実のギャップから，データを分析してもその結果が業務に活かされないということも起きてしまい，データ分析の取り組み自体の有用性に対して現場で疑念が生じることもある。

しかし，データ分析の結果に目新しさがないということは，案外自然なことである。そもそもデータをモデル化するということは，データに潜む普遍的な傾向やルールを抽象化するということである。たとえば観測データの平均値を算出するというありふれた集計も一種のモデル化であり，このモデルは統計学的には母集団の代表値を推定している。元のデータに対して説明力の高いモデルを構築しようとすれば，当然頻度の多い事象がよく反映され，結果的に経験的によく知っていることが優先的に表現される。ただ，経験的に分かっていることでも，その仮説をデータで示せたり，定性的な仮説を定量的な知識として獲得できることには十分価値があるはずであり，それが科学というものである。しかし残念ながらビジネスの現場ではそれだけでは満足されないというのも事実である。

こうした問題は PLSA でトピックを抽出するときにも該当する。PLSA ではテキストデータから観測された共起行列を教師なし学習することでトピックを抽出するが，その際の評価基準は尤度という数学的な基準であり，いかに元のデータ全体を再現できているのかということが評価される。そうして抽出されるトピックはどうしても代表的で典型的なトピックが抽出される傾向となる。そこでより特徴的で個性的なトピックを抽出し，ビジネスにおけるインサイトの獲得を狙った新しい手法として，PLSA を応用した“PCSA と“differential PLSA”という手法を開発した。

ここからはその PCSA と differential PLSA という手法の具体的な分析アプローチと，その効果を示す分析事例を紹介する。

8. 課題のターゲットに特化したトピックの抽出手法：PCSA

PCSA（Probabilistic Causal Semantic Analysis，確率的因果意味解析）[21] は，分析において特徴を見たいターゲットを定め，そのターゲットに影響を与える要因となり得るトピックを優先的に抽出する手法である。以下に手法の内容とその適用事例について説明する。

8.1 課題となるターゲットの要因の探索

ビジネスの課題解決において有効なアクションを検討するためには，その課題となるターゲットに対して要因を探ることは重要である。たとえば，商品の顧客満足度の要因を探ることで満足度の高い新商品の企画や現商品のプロモーションを検討できる。あるいは，サービスの解約や会員の退会などの要因を探ることで顧客を維持する対応やサービスの内容を検討できる。

データ分析において，こうしたターゲットの要因を探るアプローチは，回帰分析や決定木分析など，しばしば教師あり学習の手法が適用される。つまりターゲットとなる目的変数とその要因となる説明変数との関係をモデル化する。これが説明変数の候補が多量となる高次元のデータになると，教師なし学習と教師あり学習を組み合わせるアプローチがよく適用される。多量な変数をいくつかの特徴に教師なし学習で次元圧縮し，その集約された特徴を説明変数として設定し，目的変数との関係を教師あり学習でモデル化するというアプローチである。従来よく用いられてきたものとして，まず主成分分析や因子分析を実行し，その因子を説明変数に回帰分析を実行するアプローチがある。たとえば，アンケートの多量な設問項目を因子分析でいくつかの因子にまとめ，その因子と顧客満足度などのターゲット設問との関係を回帰分析して顧客の価値観を理解し，マーケティングの検討に適用されるケースが挙げられる。

先述した Nomolytics も教師なし学習と教師あり学習を組み合わせたテキストデータの分析手法と解釈できる。Nomolytics では，テキストマイニングで抽出された単語群に教師なし学習の PLSA を適用してトピックを抽出し，そのトピックを確率変数として教師あり学習のベイジアンネットワークを適用してモデルを構築している。4.3 で紹介した Nomolytics の適用例のように，たとえばアンケートデータに適用して，顧客満足度やロイヤリティの要因となる自由記述トピックを探ったり，ユーザレビューデータに適用して，レビュー評点の要因となるレビュートピックを探ったり，コールセンターの問い合わせデータに適用して，解約や退会の要因となる問い合わせトピックを探るモデルを構築できる。また，具体的な分析事例を紹介した特許文書データの分析では，技術の要因となる用途，あるいは用途の要因となる技術を探るモデルを構築した。

このように教師なし学習と教師あり学習を組み合わせることで，多量の変数を持つ高次元なビッグデータでも，次元圧縮された特徴量の変数を用いてターゲット変数との関係をシンプルに理解することができる。しかし，こうしたアプローチでは教師なし学習と教師あり学習はそれぞれ独立している。つまり，まず教師なし学習を完了してからその結果を教師あり学習に引き継ぐものであり，それは Nomolytics も同様である。そのため，教師なし学習で抽出された特徴量はどれもターゲット変数と強い関係を示すとは限らず，教師あり学習のモデリングの過程でターゲット変数に有効な特徴量のみが選択される。一方，ビジネスの課題解決の場面では，初めからターゲットに影響を与える要因に特化して特徴量が抽出される方が，効率的で望ましいとされることもある。

そこで課題となるターゲットを設定したときに，テキストデータからそのターゲットに影響を与えるトピックを優先的に抽出する手法として PCSA を開発した。以下に手法の内容を説明する。

8.2 PCSA という手法

PCSA は PLSA を応用したトピック抽出手法となる。PLSA によるトピック抽出のアプローチをおさらいすると，まずテキストデータにテキストマイニングを実行して単語を抽出し，その単語の共起頻度を集計した共起行列を作成する。そしてこの共起行列をインプットとして PLSA を適用することで，使われ方の似ている単語でまとめられたトピックを抽出する。この共起行列は全体のテキストデータから作成するため，当然そこから抽出されるトピックは全体を表現するような代表的なトピックが構成されることになる。

これに対して PCSA では，まず特徴を見たいターゲット変数を設定する。たとえば，アンケートデータでは顧客満足度，ユーザレビューデータであればレビュー評点，コールセンターの問い合わせデータでは解約希望の有無といった変数が考えられる。特許文書のデータでは，最近トレンドを形成している技術を見たければターゲット変数は出願年が候補となり，引用の多い技術を見たければターゲット変数は引用数という情報が候補となる。PCSA では，このターゲット変数を該当有無（1, 0）で示される二値変数に変換したものを使用する。

PLSA では共起行列をすべてのデータから 1 つ構築していたが，PCSA ではターゲット変数が該当するデータと該当しないデータ，それぞれから同じ行列構成の共起行列を作成する。そして，この 2 つの共起行列の差分の絶対値を計算した共起行列に対して PLSA を適用するというものである。PCSA で作成する共起行列のイメージを図 46 に示す。

全データをターゲット変数に該当するデータと該当しないデータの 2 つのグループに分割した際，それぞれのデータ件数の規模には違いがあるため，PCSA で適用する差分の共起行列はそのデータ件数の規模の違いを考慮した調整を施して差分を取る。具体的には，件数の少ないグループのデータ件数を基準に，件数の多い方のグループの共起行列の頻度をデータ件数の比率で調整する。データ件数が多い方のグループを調整対象とする理由は，その逆で調整した場合，たまたま共起したような小さな頻度が，調整によって実際には起こりえないような大きな値になってしまう可能性があるためである。つまり，件数の多い方のグループの頻度に合わせると，件数の少ない方のグループの頻度が調整され値が大きくなるが，特に 1 件などの頻度の少ない共起ペアは全体のデータ件数が多くなればその割合に応じで増加するとは言い切れない。たまたま 1 件出現したというケースは大いに考えられる。これが調整によって大きな値となると現実と乖離した共起頻度となる懸念がある。共起行列では頻度 1 件の共起ペアは非常に多く，これがすべて同じ割合で増加することの影響は大きいと考えられる。一方，件数の少ない方のグループの頻度に合わせると，件数の多い方のグループの頻度が調整され値が小さくなるが，調整後の頻度が 0 件（存在しない）になることはなく，件数の少ない方のグループの頻度を大きくすることよりも現実的と考えられる。

このようにしてターゲット変数の該当有無によって 2 つのデータに分割し，それぞれ同じ行列構成で作成された共起行列は，ターゲット変数の該当有無に影響を受ける共起ペアでは頻度の差が大きくなり，その影響を受けないような共起ペアでは頻度の差は大きく生じないことになる。つまり, 2 つの共起行列の差分を取った共起行列では, ターゲット変数の該当有無に影響を受ける共起ペアは頻度が大きくなり，そうでない共起ペアでは頻度が小さくなるということである。そのため，この共起行列に PLSA を適用することでターゲット変数の該当有無に影響を与えるような要因トピックが優先的に抽出されることが期待できる。

なお本手法 PCSA は 2023 年 2 月 6 日に特許登録されている（特許第 7221526 号）[22]。

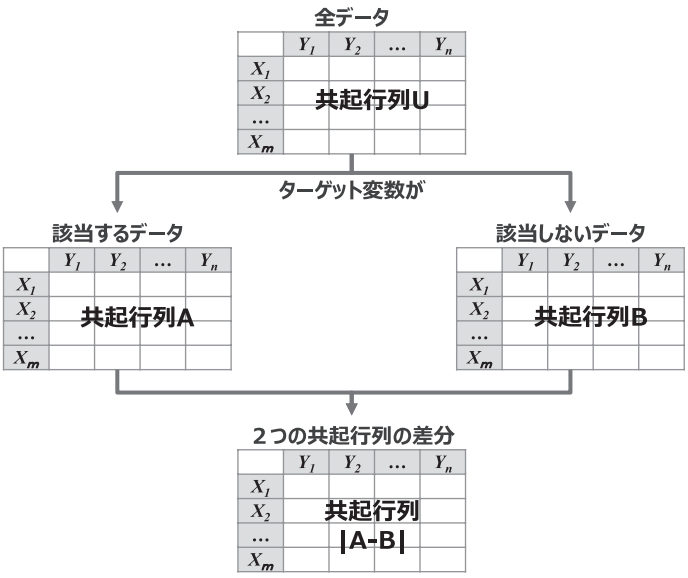


図 46　PCSA で作成する共起行列

8.3　PCSA を適用した分析事例

ここでは先述した Nomolytics の二つ目の分析事例で使用した，26,419 件の電気自動車に関連する特許文書データに PCSA を適用した分析事例を紹介し，通常の PLSA で抽出されたトピックの結果と比較する。

8.3.1　ターゲットの設定と共起行列の作成

PCSA ではまず特徴を見たいターゲット変数を設定するが，ここではそれを「出願年」とする。分析データは出願年が 2007 年から 2016 年までとなる 10 年分の特許データであるが，ターゲット変数は出願年が「2013 年以前」か「2014 年以後」という二値変数に設定する。これによって 2014 年前後で変化するような要因トピックを優先的に抽出し，トレンドが 2014 年以降で上昇傾向あるいは下降傾向にあるような技術を把握することを狙いとする。

「2013 年以前」と「2014 年以後」にデータを分割すると，それぞれのデータ件数は 20,014 件と 6,405 件となった。PCSA で作成するそれぞれの共起行列の行列構成は，全データから作成される共起行列の構成を共通して採用する。つまり，Nomolytics の分析事例で作成した「名詞単語（3,020 語）×係り受け（2,128 表現）」の共起行列の行列構成をとる。共起頻度は文章単位にカウントした頻度であるが，この共起行列の各共起ペアが 1 つでも存在する文章数は，全データでは 43,836 件であり，「2013 年以前」のデータでは 33,113 件（75.5%），「2014 年以後」のデータでは 10,723 件（24.5%）であった。「2013 年以前」のデータと「2014 年以後」のデータで，それぞれ「名詞単語（3,020 語）×係り受け（2,128 表現）」の共起行列を作成し，件数の多い「2013 年以前」のデータで作成された共起行列の頻度を各データの件数の比率でもって調整した。具体的には，「2013 年以前」のデータで作成された共起行列の全共起頻度に，文章数の比率である（10,723 / 33,113）をかけて調整済みの共起行列を作成した。このようにして作成された 2 つの共起行列の差の絶対値を計算した共起行列を作成し，これをインプットに PLSA を適用してトピックを抽出した。本分析で作成した共起行列のイメージを図 47 に示す。

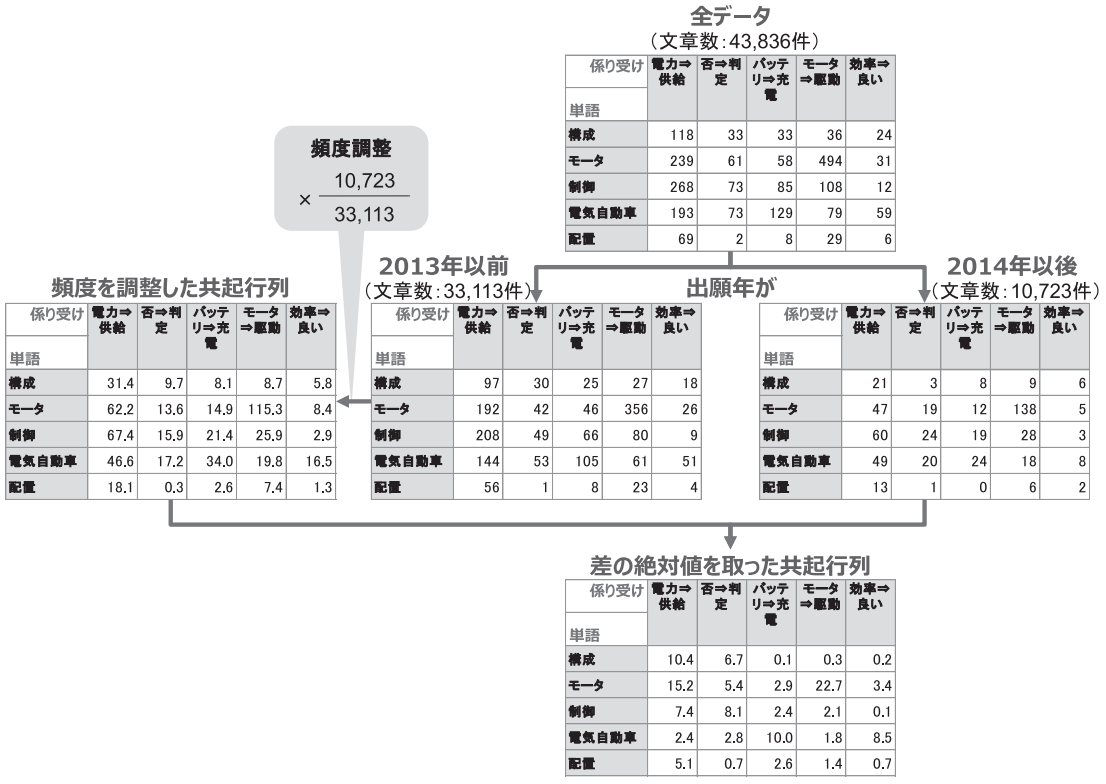


図 47　電気自動車関連の特許データに PCSA を適用するための共起行列作成

8.3.2　トピックの抽出と解釈

PLSA の実行は Nomolytics の分析事例と同様にトピック数を 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基準 AIC で評価して最も評価の良い解を採用した。PLSA の実行解の AIC の評価の結果を図 48 に示す。図 48 では，トピック数を 5 から 20 まで 1 刻みで変化させて実行した結果となっている。各トピック数に対して 5 つの黒いプロット一つひとつが初期値を変えて PLSA を実行した結果であり，灰色の折れ線グラフはその 5 つの実行解における AIC の平均値を示している。灰色の折れ線が示す AIC の平均値を比較したとき，それが最小となるトピック数は 11 個となり，このトピック数 11 個における 5 つの実行解の中で AIC が最小となる結果を採用した。

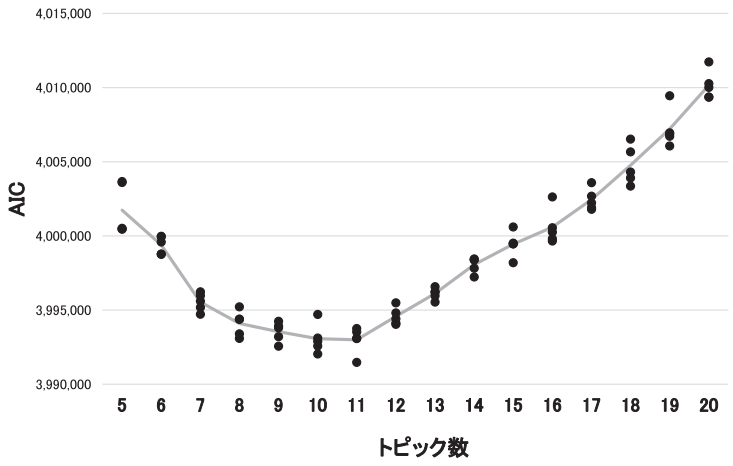


図 48　PCSA の分析における各トピック数での PLSA の実行解の AIC 評価

抽出されたトピックの内容の例を表 13 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 13（左）のトピック Zy04 では，単語は，充電，電気自動車，電力，バッテリ，蓄電装置，給電などで所属確率が高く，係り受けは，バッテリ⇒充電，充電⇒行う，電力⇒供給，電気自動車⇒充電，供給⇒電力，蓄電池⇒充電といった表現で所属確率が高い。つまり，この結果は電気自動車への充電，給電装置に関するトピックであると解釈できる。表 13(右) のトピック Zy05 では，単語は，製造，二次電池，負極，正極，製造方法，エネルギー，リチウムイオン電池などで所属確率が高く，係り受けは，水素⇒製造，発電⇒電気，製造方法⇒提供，製造⇒方法，含む⇒リチウムイオン電池といった表現で所属確率が高い。つまり，この結果は二次電池の製造方法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全 11 個のトピックの一覧を表 14 に示す。

表 13　PCSA の適用で抽出されたトピックの例

トピックZy04			
確率	単語	確率	係り受け
3.1%	充電	1.9%	バッテリ⇒充電
2.5%	電気自動車	1.9%	充電⇒行う
1.9%	電力	1.8%	電力⇒供給
1.9%	バッテリ	1.5%	電気自動車⇒充電
1.8%	蓄電装置	0.9%	供給⇒電力
1.3%	給電	0.7%	蓄電池⇒充電
1.2%	供給	0.7%	搭載⇒バッテリ
1.0%	蓄電池	0.7%	充電⇒開始
0.8%	制御	0.6%	蓄電装置⇒充電
0.8%	充電スタンド	0.6%	電気自動車⇒接続
0.7%	充電システム	0.6%	給電⇒行う
0.7%	情報	0.6%	用いる⇒充電
0.7%	充電+できる	0.5%	バッテリ⇒供給
0.7%	外部電源	0.5%	電力⇒充電
0.6%	電源	0.5%	充電⇒蓄電装置
…	…	…	…

トピックZy05			
確率	単語	確率	係り受け
3.1%	製造	4.5%	水素⇒製造
2.2%	二次電池	3.0%	発電⇒電気
1.5%	負極	2.6%	製造方法⇒提供
1.5%	正極	2.6%	製造⇒方法
1.5%	製造方法	1.6%	含む⇒リチウムイオン電池
1.2%	エネルギー	1.3%	成形品⇒提供
1.1%	リチウムイオン電池	1.3%	強度⇒有する
1.1%	水素	1.3%	表面⇒形成
1.1%	セパレータ	1.2%	含む⇒組成物
1.0%	電解液	1.2%	リレー⇒スイッチ
0.9%	形成	1.2%	二次電池⇒備える
0.9%	含有	1.2%	スイッチ⇒適する
0.9%	正極活物質	1.1%	電気部品⇒提供
0.9%	方法	1.1%	段階⇒含む
0.8%	発電	1.1%	二次電池⇒提供
…	…	…	…

表 14　PCSA の適用で抽出されたトピックの一覧

No.	トピック名
Zy01	エンジン駆動・動力伝達の制御
Zy02	モータの回転構成
Zy03	交流・直流の変換
Zy04	電気自動車への充電、給電装置
Zy05	二次電池の製造方法
Zy06	空調などの冷却・加熱
Zy07	情報通信、検出判定システム
Zy08	形成・配置
Zy09	小型化・低コスト化・簡素化・操作性向上
Zy10	重力発電の活用による地球温暖化防止
Zy11	既存エンジンへの警鐘、タービン発電・重力発電

8.3.3　トピックのスコアリング

Nomolytics の分析事例と同様に，分析対象とした 26,419 件の特許データに対して，今回抽出された 11 個のトピックのスコア（該当度）を計算した。スコアの計算方法は Nomolytics の分析事例と同様であるため割愛するが，連続値であるスコアから該当有無を示す 1,0 のフラグに変換する閾値は，スコアの分布や実際の文章の内容も確認しながら 5 に設定した。トピックのスコアリングの計算処理により，表 15 に示すようなデータが作成された。26,419 件の特許データに対して，Nomolytics の分析事例で全体のデータから抽出された 34 個のトピックに加え，今回の PCSA の適用によって抽出された 11 個のトピックの 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このトピックを軸にしたさまざまな分析を実行することができる。

表 15　PCSA で抽出したトピックの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	出願年グループ	出願人	要約	トピックZ01	トピックZ02	…	トピックZ34	トピックZy01	トピックZy02	…	トピックZy11
1	特願2007-XX	2007	①2013年以前	A社	【課題】電気式変速	1	1		0	0	1		1
2	特願2009-XX	2009	①2013年以前	B社	【課題】従来の電気	0	1		1	0	0		1
3	特願2012-XX	2012	①2013年以前	C社	エンジンのための	0	1		1	0	1		0
4	特願2013-XX	2013	②2014年以後	D社	【課題】駐車場に設	1	0		0	1	0		0
…	…	…	…	…		…	…		…	…	…		…
26,419	特願2016-XX	2016	②2014年以後	X社	充電ステーションが	1	0		1	1	0		1

8.3.4　PLSA と PCSA の結果の比較

本分析では，ターゲット変数を「出願年」に設定し，全体で 26,419 件の特許データを「2013 年以前」と「2014 年以後」の 2 つのグループに分割して PCSA を適用した。分析の狙いは，出願年が 2014 年前後で変化するような要因トピックを優先的に抽出し，トレンドが 2014 年以降で上昇傾向あるいは下降傾向にあるような技術を把握することだった。実際に抽出された 11 個のトピックにそうした傾向があるのか，Nomolytics の分析事例で全体のデータから抽出された 34 個のトピックと比較して確認した。

確認の仕方は，表 15 のデータを用いて，Nomolytics の PLSA で全体のデータから抽出された 34 個のトピックと，今回の PCSA で抽出された 11 個のトピックについて，それぞれのトピックが該当する特許のうち，出願年が「2014 年以後」となっている割合を比較した。これによって，2014 年前後において PCSA ではどれくらい偏ったトピックを抽出できているか確認した。その結果を図 49 に示す。図 49 のグラフは，横軸をトピック，縦軸を出願年が 2014 年以後となっている割合とし，その割合が高い順に横軸のトピックを並べている。図 49 の上のグラフは Nomolytics の PLSA で全体のデータから抽出された 34 個のトピックの結果であり，下のグラフは PCSA で抽出された 11 個のトピックの結果である。この上下のグラフを比較すると，割合の平均値はどちらもほぼ同じだが，そのば

らつきには大きな違いが生まれている。PLSA で抽出された全体トピック 34 個の割合の大きさはおおむね 25％前後となっているが，PCSA で抽出された要因トピック 11 個は，割合の大きさが高いものと低いものに偏っている。つまり，PCSA では出願年が 2014 年前後で特徴を示すトピックが抽出されていることが確認できる。

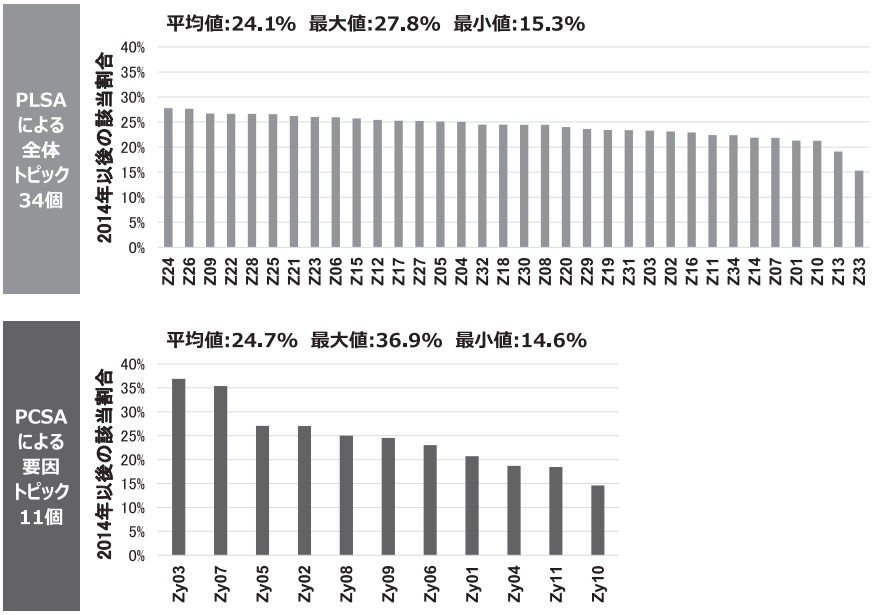


図 49　全体のトピック 34 個と PCSA のトピック 11 個の比較

8.3.5　トレンドが上昇傾向にあるトピックの確認

PCSA で抽出されたトピックのうち，図 49（下）で 2014 年以後の割合が高かった上位 2 つのトピック Zy03 と Zy07 について，実際のトレンドを確認した。Nomolytics の分析事例で紹介したトレンド分析と同様に，出願年とトピックの関係の強さを示すリフト値を使ってそれぞれの経年変化を可視化した。その結果を図 50 に示す。図 50 より，確かにどちらもきれいな上昇トレンドを形成していることがわかる。なお，Zy03 は「交流・直流の変換」に関するトピック，Zy07 は「情報通信，検出判定システム」に関するトピックであるが，それぞれのトピックの構成内容と，該当する特許の要約の一例を図 51, 52 に示す。たとえば，Zy03 は，電気自動車が省電力動作をとるときには，コンバータが直流電力を通常よりも低い電圧に変換するが，冷却ポンプへは高い効率で電力を供給することで，十分な冷却性能を得るという技術がある。Zy07 は，電気自動車の圧力チャンバに設置されている圧力センサで算出される圧力値と，外周面の導体版から形成されるコンデンサで算出される電流値により，車両への衝突を検知する技術がある。

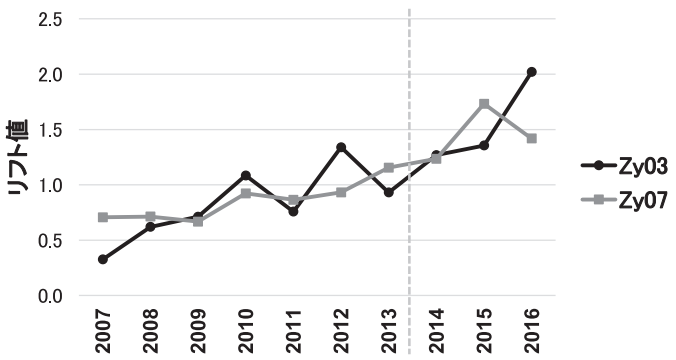


図 50　2014 年以後の割合が高いトピック Zy03 と Zy07 のトレンド

トピックの構成

トピックZy03

確率	単語	確率	係り受け
1.9%	供給	1.3%	交流電力−変換
1.5%	変換	1.2%	直流電力−変換
1.5%	制御	1.0%	電力−供給
1.3%	電圧	0.9%	変換−供給
1.2%	インバータ	0.8%	直列−接続
1.2%	電力	0.8%	並列−接続
1.1%	制御部	0.6%	供給−電力
1.1%	検出	0.6%	バッテリ−接続
1.1%	バッテリ	0.6%	電力−変換
1.1%	直流電力	0.5%	モータ−供給
1.1%	電気自動車	0.5%	電圧−検出
1.0%	モータ	0.5%	交流−変換
1.0%	交流電力	0.5%	直流−変換
0.9%	スイッチ	0.5%	負荷−供給
0.8%	直流	0.5%	直流電圧−変換
…	…	…	…

該当する特許の一例

発明の名称	出願年
電気自動車用の電源システム	2016

要約文

通常動作と省電力動作の何れにおいても，十分な冷却性能が得るとともに，ポンプを高効率で動作させる電源システムを提供する。コンバータが，変換後の直流電力の電圧を第1電圧に制御する通常動作と，第1電圧よりも低い第2電圧に制御する省電力動作を実行可能である。電力制御ユニットを冷却する冷媒を循環させるポンプの制御ユニットが，省電力動作時に通常動作時よりも変換後の直流電力がポンプに供給される期間の比率を高くする。

※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 51　トピック Zy03「交流・直流の変換」の構成内容と特許の一例

トピックの構成

トピックZy07

確率	単語	確率	係り受け
1.8%	検出	1.3%	否−判定
1.4%	方法	0.7%	システム−備える
1.3%	判定	0.6%	基づく−検出
1.3%	制御	0.6%	基づく−生成
1.1%	構成	0.6%	方法−備える
1.1%	受信	0.6%	電気信号−変換
1.0%	演算	0.5%	信号−受信
1.0%	制御部	0.5%	制御部−有する
0.9%	生成	0.5%	車両−走行
0.9%	信号	0.4%	電気信号−出力
0.8%	システム	0.4%	情報−基づく
0.8%	センサ	0.4%	制御−構成
0.7%	変化	0.4%	有無−判定
0.7%	測定	0.4%	制御部−含む
0.7%	制御装置	0.4%	モータ−制御
…	…	…	…

該当する特許の一例

発明の名称	出願年
車両用衝突検知装置	2015

要約文

車両への衝突が発生したことを精度よく検出することができる車両用衝突検知装置の提供。バンパーアブソーバーには，圧力チャンバが配置されている。圧力チャンバには圧力センサが接続されている。圧力チャンバの外周面に貼付されている前方導体板および後方導体板により，複数のコンデンサが形成されている。圧力検出値から算出された有効質量が閾値以上であって，かつ，コンデンサの容量の増大により，電流検出値が閾値以上である時に，車両への衝突を検出する。

※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 52　トピック Zy07「情報通信，検出判定システム」の構成内容と特許の一例

8.3.6　PCSA のまとめ

PCSA は，特徴を見たいターゲットを定め，そのターゲットに影響を与える要因となり得るトピックを優先的に抽出する手法である。以下に PCSA の手法の特徴を Nomolytics と比較してまとめる。

Nomolytics では，まずテキストデータ全体を表す代表的なトピックを PLSA で抽出し，そのトピックの特徴を，たとえば出願年や出願人といったさまざまな分析軸で探索することができる。一方，PCSA は，たとえば最近トレンドを形成している技術を把握したいというように，特徴を探索したいターゲットが定まっているときに適用する手法となる。そのターゲットに特化したトピックを優先的に抽出し，より顕著な傾向を示す要因を深く分析することでインサイトの獲得を狙う手法である。

どちらが優れているというものではなくて，目的に応じて使い分けることが重要である。PCSA は分析軸を最初から一つに絞り，それに特化したトピックを抽出できるが，Nomolytics のようにさまざまな分析軸で広く特徴を探索したいというときには不向きとなる。ビジネスの課題解決において，特定された課題の要因を探り，有効な施策アイデアを効率的に検討したい場合は，PCSA を適用することで有用な知識が得られるものと期待できる。

9. 個性的なトピックの抽出手法：differential PLSA

differential PLSA [23] は，より個性的なトピックを抽出することでインサイトの獲得を狙う手法となる。以下に手法の内容とその適用事例について説明する。

9.1 典型的でない個性的なトピックの抽出

第 7 項で述べたとおり，ビジネスにおけるデータ活用の場面では，新たな気づきとなるインサイトの獲得が期待されるものの，実際にデータ分析から得られる結果は，経験的によく知っているものも多く，目新しさがないと評価されてしまうことがある。テキストデータから PLSA でトピックを抽出する分析でも，そのトピックはデータ全体を代表するような典型的なトピックが抽出される傾向がある。次元圧縮法である PLSA は，元の高次元のデータに対して，再現性の高い低次元のトピックに変換する手法であるため，元のデータに対して表現力の高いトピックが抽出される。PLSA のインプットとなる共起行列は頻度を値とするデータであるが, 元のデータに対して表現力の高いトピックを抽出するには，当然頻度の高い要素を中心に代表的なトピックが形成される傾向がある。結果として高い頻度の要素で構成され，経験的によく知っているような，典型的とも感じられるトピックが抽出されるのである。しかし，これはデータ全体の特徴を把握する上では有用な分析であり，一つひとつのトピックはよく知っている典型的なものであっても，そうしたトピックに類型化することは人間では困難な処理のはずである。実際に結果に対して目新しさがないと感じる人は，結果を見たからそう感じているだけで，結果を見ないで自分の経験や感覚に基づいて代表的なトピックを網羅的に列挙することはほぼできない。

ここでは，典型的であってもデータ全体を表現するトピックを抽出することで，全体像を把握する分析の重要性を前提とした上で，とは言うものの，より個性的なトピックも抽出できる手法として開発した differential PLSA について説明する。

9.2 differential PLSA という手法

differential PLSA も PCSA のように，共起行列を加工して PLSA を適用する手法となる。differential PLSA では，行要素 Xi と列要素 Yj が共起する実測頻度 n（Xi,Yj）を値として持つ通常の共起行列 M に加え, 行要素 Xi と列要素 Yj が共起する期待頻度 n’（Xi,Yj ）を値として持つ共起行列 M’ を作成する。期待頻度とは，行要素 Xi と列要素 Yj の総頻度（出現文章数）と全体における出現割合に基づいて計算した理論値となる。具体的には，Xi の総頻度（出現文章数）を n（Xi），Yj の総頻度（出現文章数）を n（Yj），総文章数を N とすると，行要素 Xi の総頻度 n（Xi）に対して，列要素 Yj の出現割合 n（Yj）/N をかけた値となり，以下の式（15）のように計算される。あるいは，列要素 Yj の総頻度 n（Yj）に対して，行要素 Xi の出現割合 n（Xi）/N をかけた値でも同じである。この実測頻度と期待頻度の 2 つの共起行列は，カイ二乗検定をするときに作成する観測度数と期待度数の 2 つのクロス集計表に近いイメージのものである。differential PLSA は，共起行列の各共起ペアにおいて，期待頻度に対する実測頻度の比率の対数を取った log（n(Xi, Yj)/n′(Xi, Yj)）を値として持つ共起行列（differential 共起行列）を構築し, これに PLSA を適用する。なお，対数の計算において値が負数となるものは 0 に置換する。differential PLSA で作成する共起行列のイメージを図 53 に示す。

$$n'(Xi,Yj) = n(Xi) \cdot n(Yj) / N \qquad (15)$$

実測頻度の共起行列に適用する通常の PLSA では，その解を求める最適化計算において，どうしても頻度が高い要素に高い確率が割り当てられ，結果として抽出されるトピックは典型的なものになる傾向があり，目新しさに欠けてしまう。一方，differential PLSA の共起行列では，実測頻度を期待頻度で除した値を持つが，実測頻度が高い共起ペアでも，元々全体の頻度が高い要素が含まれるときには期待頻度も高い値となるため，実測頻度を期待頻度で除すことで値の大きさが制限される。逆に実測頻度が高くない共起ペアでも，期待頻度がそれよりも十分小さければ値は大きくなり，これに PLSA を適用した解ではこうした要素にも高い確率が割り当てられる可能性がある。つまり，通常

の PLSA では頻度が低い要素には高い確率が割り当てられにくい傾向があるが，differential PLSA ではそうした要素にも高い確率が割り当てられる可能性があり，より個性的なトピックが抽出されることが期待できる。

また，期待頻度に対する実測頻度の比率に対数を取る理由は，極端に高くなる値を制限するためである。特に期待頻度は 1 未満となるケースも多く，比率のみでは値が高くなりすぎるものもある。この状態では共起行列全体の値の分布は大きくばらつき，極端な値の開きが生まれてしまう。このまま PLSA を適用した場合，この極端に大きな値に引っ張られる結果となるため，必要以上にデフォルメされた歪んだトピックとなることが考えられる。そこでこの比率の値の対数を取ることで値の分布をならし，この問題を緩和する。

なお本手法 differential PLSA は 2023 年 2 月 6 日に特許登録されている（特許第 7221527 号）[24]。

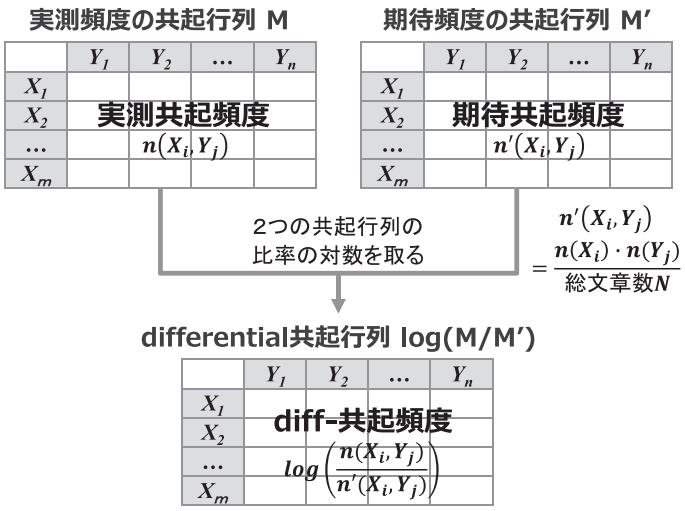


図 53　differential PLSA で作成する共起行列

9.3　differential PLSA を適用した分析事例

ここでは先述した Nomolytics の二つ目の分析事例で使用した，26,419 件の電気自動車に関連する特許文書データに differential PLSA を適用した分析事例を紹介し，通常の PLSA で抽出されたトピックの結果と比較する。

9.3.1　共起行列の作成

differential PLSA では実測頻度の共起行列と期待頻度の共起行列を作成するが，実測頻度の共起行列は Nomolytics の分析事例で作成した「名詞単語（3,020 語）×係り受け（2,128 表現）」の共起行列となる。期待頻度の共起行列は，これと同じ行列構成をとり，各共起ペアの期待頻度を計算する。なお，今回のデータにおいて，総文章数 N は 229,598 件であった。この総文章数とは，テキストマイニングの実行の際に，文章の区切り文字に指定した句点（。）や隅付括弧（【 】）によって分割された文章の数であり，共起行列を構成する単語や係り受けが一切登場しない文章も含まれる。共起行列の各共起ペアにおいて，期待頻度に対する実測頻度の比率の対数を取った共起行列を作成し，これをインプットに PLSA を適用することでトピックを抽出した。本分析で作成した共起行列のイメージを図 54 に示す。

実測頻度の共起行列

全体の頻度	係り受け / 単語	電力⇒供給	否⇒判定	バッテリ⇒充電	モータ⇒駆動	効率⇒良い
全体の頻度		1,350	539	529	494	444
5,880	構成	118	33	33	36	24
5,188	モータ	239	61	58	494	31
5,092	制御	268	73	85	108	12
4,655	電気自動車	193	73	129	79	59
4,604	配置	69	2	8	29	6

期待頻度の共起行列

全体の頻度	係り受け / 単語	電力⇒供給	否⇒判定	バッテリ⇒充電	モータ⇒駆動	効率⇒良い
全体の頻度		1,350	539	529	494	444
5,880	構成	34.6	13.8	13.5	12.7	11.4
5,188	モータ	30.5	12.2	12.0	11.2	10.0
5,092	制御	29.9	12.0	11.7	11.0	9.8
4,655	電気自動車	27.4	10.9	10.7	10.0	9.0
4,604	配置	27.1	10.8	10.6	9.9	8.9

differential共起行列

全体の頻度	係り受け / 単語	電力⇒供給	否⇒判定	バッテリ⇒充電	モータ⇒駆動	効率⇒良い
全体の頻度		1,350	539	529	494	444
5,880	構成	1.2	0.9	0.9	1.0	0.7
5,188	モータ	2.1	1.6	1.6	3.8	1.1
5,092	制御	2.2	1.8	2.0	2.3	0.2
4,655	電気自動車	2.0	1.9	2.5	2.1	1.9
4,604	配置	0.9	0.0	0.0	1.1	0.0

図 54　電気自動車関連の特許データに differential PLSA を適用するための共起行列作成

9.3.2　トピックの抽出と解釈

PLSA の実行は Nomolytics の分析事例と同様にトピック数を 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基準 AIC で評価して最も評価の良い解を採用した。PLSA の実行解の AIC の評価の結果を図 55 に示す。図 55 では，トピック数を 35 から 60 まで 1 刻みで変化させて実行した結果となっている。各トピック数に対して 5 つの黒いプロット一つひとつが初期値を変えて PLSA を実行した結果であり，灰色の折れ線グラフはその 5 つの実行解における AIC の平均値を示している。灰色の折れ線が示す AIC の平均値を比較したとき，それが最小となるトピック数は 50 個となり，このトピック数 50 個における 5 つの実行解の中で AIC が最小となる結果を採用した。解釈を付けた 50 個のトピックの一覧を表 16 に示す。

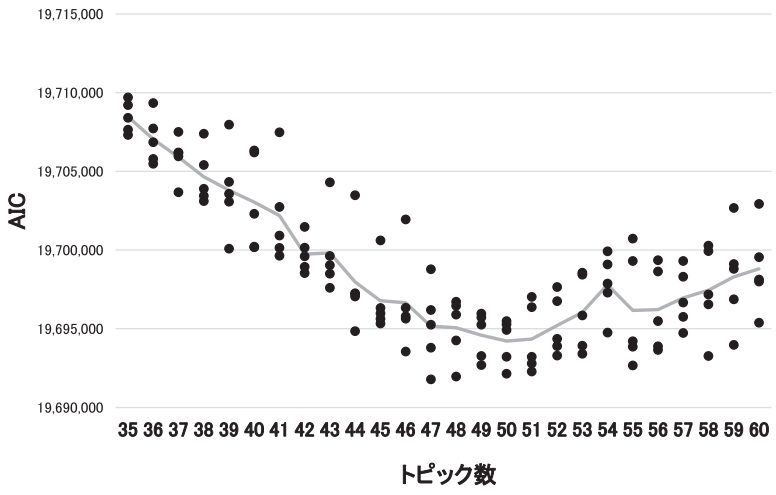


図 55　differential PLSA の分析における各トピック数での PLSA の実行解の AIC 評価

表 16　differential PLSA の適用で抽出されたトピックの一覧

No.	トピック名	No.	トピック名
Z'01	エンジン制御	Z'26	電流・電圧の検出
Z'02	動力伝達	Z'27	温度，電流，充電量などの検出と制御
Z'03	差動機構などを備えた動力伝達の制御	Z'28	演算や推定，測定などのステップを含む方法
Z'04	回転運動	Z'29	情報の取得・提供（位置情報やバッテリ残量等）
Z'05	ロータ・ステータなどモータの構成	Z'30	スイッチなど操作装置
Z'06	モータ制御（トルク制御や回転数制御等）	Z'31	車両用灯具
Z'07	油圧ポンプなどを利用したモータ駆動	Z'32	掃除機
Z'08	ブレーキ	Z'33	基板の構成
Z'09	状態に応じた制御，運転者の操作補助	Z'34	回路の接続（電力変換回路等）
Z'10	コンバータとバッテリ昇降圧	Z'35	端子接続
Z'11	直流と交流の電力変換	Z'36	部品・装置の収容ケース・筐体
Z'12	回転力などの電気エネルギー変換	Z'37	部品・装置の配置
Z'13	エネルギー効率の向上	Z'38	パーツなどの移動，位置
Z'14	発電と蓄電	Z'39	構造の形成・方位
Z'15	電池モジュールの提供	Z'40	支持構造
Z'16	燃料電池	Z'41	装置やユニットの構成
Z'17	二次電池の構成	Z'42	システム・方法の構成
Z'18	バッテリの充放電	Z'43	その他方法
Z'19	充電システム	Z'44	組成物の製造方法（樹脂や電解液等）
Z'20	充電の接続	Z'45	機能性組成物・成形品（耐熱性や耐衝撃性等）
Z'21	非接触など受給電装置	Z'46	製造の効率化（小型化や低コスト化等）
Z'22	車両用空調など熱交換	Z'47	不具合の防止（損傷，感電，盗難等）
Z'23	冷却装置と放熱	Z'48	その他
Z'24	信号の入出力と検出	Z'49	タービン発電と船舶・飛行機への応用
Z'25	電気信号の取得と変換（センサ検出等）	Z'50	重力発電の活用による地球温暖化防止

9.3.3　トピックのスコアリング

Nomolytics の分析事例と同様に，分析対象とした 26,419 件の特許データに対して，今回抽出された 50 個のトピックのスコア（該当度）を計算した。スコアの計算方法は Nomolytics の分析事例と同様であるため割愛する。連続値であるスコアから該当有無を示す 1,0 のフラグに変換する閾値は，スコアの分布や実際の文章の内容も確認しながら 10 に設定した。トピックのスコアリングの計算処理により，表 17 に示すようなデータが作成された。26,419 件の特許データに対して，Nomolytics の分析事例で全体のデータから抽出された 34 個のトピックに加え，今回の differential PLSA の適用によって抽出された 50 個のトピックの 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このトピックを軸にしたさまざまな分析を実行することができる。

表 17　differential PLSA で抽出したトピックの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	トピック Z01	トピック Z02	…	トピック Z34	トピック Z'01	トピック Z'02	…	トピック Z'50
1	特願2007-XXX	2007	A社	【課題】電気式変速操	1	1		0	0	0		1
2	特願2009-XXX	2009	B社	【課題】従来の電気自	0	1		1	1	0		0
3	特願2012-XXX	2012	C社	エンジンのための方法	0	1		1	0	1		1
4	特願2013-XXX	2013	D社	【課題】駐車場に設置	1	0		0	1	0		0
…	…	…	…		…	…		…	…	…		…
26,419	特願2016-XXX	2016	X社	充電ステーションが電	1	0		1	0	1		0

9.3.4　PLSA と differential PLSA の結果の比較

Nomolytics で適用する通常の PLSA では，データ全体を表現する典型的なトピックが抽出される傾向があることに対して，differential PLSA は，より個性的なトピックを抽出できる手法として開発したものである。実際に differential PLSA で抽出された 50 個のトピックはそうした傾向があるのか，Nomolytics の分析事例において通常の PLSA によって抽出された 34 個のトピックと比較し，その違いを確認した。注目すべき違いとして，(1) differential PLSA は頻度の低い要素でも高い確率が割り当てられること，(2) differential PLSA でのみ抽出されるトピックがあること，(3) differential PLSA の方がトピックの解釈が難しいことが挙げられる。以下にそれぞれについて結果を交えながら説明する。

(1) differential PLSA は頻度の低い要素でも高い確率が割り当てられる

この特徴に関しては，通常の PLSA と differential PLSA で同様の解釈ができるトピックの内容を並べて比較すると分かりやすい。

たとえば，「ブレーキ」に関するトピックでは，通常の PLSA では「Z05. ブレーキ装置」として，differential PLSA では「Z' 08. ブレーキ」として抽出されているが，それぞれのトピックの構成内容を比較したものを表 18 に示す。表 18 では，それぞれのトピックにおいて単語と係り受けを所属確率の高い順に並べているが，確率に加えてその単語・係り受けの総頻度（出現文章数）も掲載している。表 18 の比較より，表 18（左）の通常の PLSA によるトピック Z05 は，所属確率の高い単語は全体の出現頻度も高い傾向にあるが，表 18（右）の differential PLSA によるトピック Z' 08 は，全体の出現頻度が低い単語でも高い所属確率が割り当てられている。そのため，通常の PLSA によるトピック Z05 では，「ブレーキ」や「制動力」といった，ブレーキの直接的な表現が上位に位置しており，まさにブレーキを代表するトピックとなっている。一方，differential PLSA によるトピック Z' 08 では，「マスタシリンダ」や「液圧」といった，頻度は低いがより具体的なブレーキ装置の中身に関する表現が上位にある。

もう一つの比較の例として，電気自動車の「非接触などの給電」に関するトピックでは，通常の PLSA では「Z14. 非接触受電などの給電装置」として，differential PLSA では「Z' 21. 非接触など受給電装置」として抽出されている。それぞれのトピックの構成内容を比較したものを表 19 に示す。表 19 の比較より，表 19（左）の通常の PLSA によるトピック Z14 は，やはり所属確率の高い単語は全体の出現頻度も高い傾向にあるが，表 19（右）の differential PLSA によるトピック Z' 21 は，全体の出現頻度が低い単語でも高い所属確率が割り当てられている。そのため，通常の PLSA によるトピック Z14 では，「給電」「電力」「電源」といった，給電の直接的な表現ばかりが上位に位置しており，まさに給電を代表するトピックとなっている。一方，differential PLSA によるトピック Z' 21 では，「非接触」という単語が最も確率が高く，また「送電コイル」や「受電コイル」といった，頻度は低いがより具体的な非接触給電の装置に関する表現が上位にある。電気自動車において，非接触の充電システムは今後重要な技術と考えられるが，その仕組みは，駐車場の路面に設置した送電コイルと車両に搭載された受電コイルが重なり合うことで電磁誘導が発生し，電力が供給されるという仕組みが代表的である。differential PLSA ではそうした技術のより具体的なキーワードがトピックの上位語となっている。

各トピックにおいて所属確率の高い上位語の頻度が，通常の PLSA よりも differential PLSA の方が少ないことを確認するため統計的検定も行った。その方法は，各トピックにおいて所属確率の高い順に単語および係り受けを並べたときに，その累積確率が 50%になるまでの単語および係り受けの平均頻度を検定用データとし，通常の PLSA で抽出された 34 個のトピックと differential PLSA で抽出されたトピック 50 個のトピックの平均値の差を Welch の t 検定で検定した。通常の PLSA では単語の頻度は平均 1288.2 件（SD = 686.5），係り受けの頻度は平均 85.6 件（SD = 26.0），differential PLSA では単語の頻度は平均 440.9 件（SD = 137.5），係り受けの頻度は平均 58.8 件（SD = 15.9）となり，単語および係り受けの両方で 1%有意の違いがみられた。

表 18 「ブレーキ」に関するトピックにおける通常の PLSA と differential PLSA の比較

通常のPLSAで抽出されたトピックZ05					
確率	頻度	単語	確率	頻度	係り受け
7.3%	779	ブレーキ	3.2%	92	車両⇒ブレーキ
5.0%	1,384	作動	2.4%	33	ブレーキ液圧⇒発生
3.4%	5,188	モータ	2.2%	53	制動力⇒発生
2.9%	279	制動力	1.8%	49	ブレーキ⇒備える
2.9%	867	運転者	1.7%	32	操作量⇒応ずる
2.7%	1,117	車輪	1.6%	82	ブレーキ⇒提供
2.6%	1,005	操作	1.6%	59	電気信号⇒基づく
1.6%	111	操作量	1.6%	41	運転者⇒操作
1.6%	162	ブレーキペダル	1.6%	55	操作⇒応ずる
1.5%	5,092	制御	1.5%	235	モータ⇒制御
1.4%	485	電気信号	1.4%	32	基づく⇒発生
1.2%	73	ブレーキ液圧	1.3%	494	モータ⇒駆動
1.2%	299	付与	1.3%	34	制動トルク⇒発生
1.2%	3,960	検出	1.3%	21	液圧⇒発生
1.1%	117	液圧	1.3%	22	ブレーキペダル⇒操作
…		…	…		…

differential PLSAで抽出されたトピックZ'08					
確率	頻度	単語	確率	頻度	係り受け
2.2%	112	マスタシリンダ	3.2%	32	基づく⇒発生
2.0%	73	ブレーキ液圧	2.6%	32	操作量⇒応ずる
1.6%	72	ブレーキ操作	2.6%	33	ブレーキ液圧⇒発生
1.6%	117	液圧	2.5%	53	制動力⇒発生
1.4%	779	ブレーキ	2.3%	49	ブレーキ⇒備える
1.4%	279	制動力	2.1%	20	備える⇒ブレーキ
1.3%	111	操作量	2.0%	92	車両⇒ブレーキ
1.3%	162	ブレーキペダル	2.0%	22	ブレーキペダル⇒操作
1.1%	122	電気ブレーキ	1.7%	21	液圧⇒発生
1.0%	38	スレーブシリンダ	1.7%	55	操作⇒応ずる
0.9%	25	モータシリンダ装置	1.6%	41	運転者⇒操作
0.9%	299	付与	1.5%	32	応ずる⇒発生
0.9%	43	制動操作部材	1.4%	28	ブレーキ⇒有する
0.9%	39	ブレーキ液	1.3%	25	応ずる⇒電気信号
0.9%	65	摩擦部材	1.3%	26	制動力⇒付与
…		…	…		…

表 19 「非接触などの給電」に関するトピックにおける通常の PLSA と differential PLSA の比較

通常のPLSAで抽出されたトピックZ14					
確率	頻度	単語	確率	頻度	係り受け
9.7%	1,162	給電	4.3%	1,350	電力⇒供給
5.1%	3,629	電力	2.7%	115	給電⇒行う
3.6%	1,140	電源	2.5%	52	電力⇒受電
3.4%	329	給電装置	2.0%	28	給電⇒電力
2.4%	4,655	電気自動車	1.6%	124	電源⇒接続
2.4%	180	非接触	1.5%	20	駐車装置⇒変化
2.4%	1,052	外部	1.5%	34	非接触⇒受電
2.3%	160	受電	1.3%	85	給電⇒停止
1.3%	158	駐車	1.3%	35	受電⇒電力
1.2%	398	電源装置	1.3%	70	給電⇒制御
1.2%	5,092	制御	1.2%	40	給電装置⇒備える
1.2%	124	受電部	1.2%	57	給電⇒受ける
1.2%	1,064	変化	1.1%	25	電力⇒給電
1.2%	63	駐車装置	1.1%	27	電力⇒送電
1.1%	3,978	供給	1.1%	41	車両⇒給電
…		…	…		…

differential PLSAで抽出されたトピックZ'21					
確率	頻度	単語	確率	頻度	係り受け
3.2%	180	非接触	3.0%	52	電力⇒受電
2.5%	78	送電コイル	2.7%	28	給電⇒電力
2.4%	160	受電	2.3%	35	受電⇒電力
2.4%	86	受電コイル	2.1%	25	電力⇒給電
2.2%	329	給電装置	1.8%	20	駐車装置⇒変化
2.0%	73	受電装置	1.7%	34	非接触⇒受電
1.8%	124	受電部	1.6%	21	供給⇒給電装置
1.6%	63	駐車装置	1.6%	27	電力⇒送電
1.6%	130	給電+できる	1.4%	40	給電装置⇒備える
1.3%	36	電磁誘導	1.1%	115	給電⇒行う
1.3%	105	送電	1.1%	41	車両⇒給電
1.2%	95	パレット	0.9%	39	電力⇒伝送
1.2%	41	給電回路	0.9%	47	電力⇒受ける
1.2%	1,162	給電	0.7%	45	外部⇒供給
1.1%	83	駐車スペース	0.7%	49	車両⇒移動
…		…	…		…

（2）differential PLSA でのみ抽出されるトピックがある

differential PLSA では通常の PLSA では抽出されないトピックが抽出されていた。これは，differential PLSA の方が抽出トピック数が多かったということもあるが，通常の PLSA のトピック数を 50 個に設定して計算した解も確認したうえで，differential PLSA のみに現れたトピックが存在していた。その例を図 56, 57 に示す。図 56, 57 はそれぞれ differential PLSA でのみに抽出されたトピック Z’ 09 と Z’ 29 であり，そのトピックの構成内容と該当する特許の要約文の一例を掲載したものとなる。図 56 で示すトピック Z’ 09 は，単語では，シフトレンジ，パーキングレンジ，検出結果，停止などで所属確率が高く，係り受けでは，自動的⇒行う，操作⇒行う，駆動⇒停止といった表現で所属確率が高い。これは運転者の操作を補助したり，自動停止などの運転アシストに関する技術と解釈できる。図 57 で示すトピック Z’ 29 は，単語では，ナビゲーション装置，情報，目的地，位置情報などで所属確率が高く，係り受けでは，情報⇒送信，情報⇒含む，情報⇒取得，情報⇒受信といった表現で所属確率が高い。これは位置情報を取得してドライバーにナビ情報として提供するといった，情報の獲得と提供に関する技術と解釈できる。

こうした技術は，頻度の高い要素を中心に代表的なトピックを抽出する通常の PLSA では抽出しきれなかった情報であるが，どちらも自動車の付加価値を高める重要な技術といえる。

トピックZ'09					
確率	頻度	単語	確率	頻度	係り受け
1.4%	50	シフトレンジ	1.9%	23	自動的⇒行う
1.0%	38	パーキングレンジ	1.7%	38	操作⇒行う
0.9%	147	検出結果	1.6%	22	駆動⇒停止
0.7%	1,200	停止	1.6%	26	動作⇒行う
0.7%	365	解除	1.6%	40	要する⇒時間
0.6%	34	キースイッチ	1.4%	46	停止⇒状態
0.6%	3,960	検出	1.2%	26	ブレーキ⇒作動
0.6%	237	禁止	1.1%	96	状態⇒検出
0.5%	204	切替	1.1%	62	状態⇒切り替える
0.5%	67	閉状態	1.1%	35	発生⇒検出
0.5%	1,005	操作	1.1%	57	制御装置⇒行う
0.5%	199	移行	1.1%	23	作動⇒行う
0.5%	1,384	作動	1.1%	22	介する⇒制御
0.5%	62	絶縁抵抗	1.0%	25	モータ⇒停止
0.5%	93	シフトレバー	1.0%	40	ロック⇒解除
…		…	…		…

該当する特許の要約文の一例

【課題】モータによる長期間のクリープトルク発生による電力消費を抑えるよう促すとともに，車両停止時の駆動源の切り替りによる車両の動き出しの防止又は車両の動き出しを運転者に知らせることのできるハイブリッド電気自動車を提供すること。【解決手段】アクセルペダル及びブレーキペダルの踏み込みがなく，エンジンの作動中であり，シフト位置が走行レンジ（Dレンジ又はRレンジ）である車両停止時に，クリープ走行制御の駆動源がモータからエンジンに切り替わった際のトルク変動により車両が動き出した際には，パーキングブレーキによる制動を自動的に行い，且つ警報器及び警告灯による運転者への注意喚起を行う。

※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る

図 56　differential PLSA でのみ抽出された「運転者の操作補助」に関するトピック

トピックZ'29					
確率	頻度	単語	確率	頻度	係り受け
1.2%	122	ナビゲーション装置	2.1%	48	情報⇒送信
1.1%	809	情報	2.1%	42	情報⇒含む
1.0%	165	目的地	1.8%	68	情報⇒取得
1.0%	111	位置情報	1.5%	31	情報⇒受信
0.9%	786	取得	1.5%	35	示す⇒情報
0.9%	819	送信	1.5%	126	情報⇒基づく
0.8%	558	表示	1.4%	25	情報⇒用いる
0.8%	43	検索	1.3%	48	表示部⇒表示
0.8%	72	バッテリ残量	1.2%	31	基づく⇒特定
0.8%	83	サーバ	1.2%	42	無線通信⇒行う
0.7%	511	記憶	1.2%	29	情報⇒提供
0.7%	877	受信	1.1%	28	記憶⇒記憶手段
0.7%	83	無線通信	1.1%	29	基づく⇒取得
0.7%	61	現在地	1.0%	32	車両⇒位置
0.7%	51	地図情報	1.0%	33	情報⇒記憶
…		…	…		…

該当する特許の要約文の一例

【課題】充電スタンドへの電気自動車の有効なナビゲーションを行う。【解決手段】本発明は，複数の充電スタンドの位置情報を記憶するデータベースと，充電スタンドが保持している電池についてその種別およびエネルギー残量を記憶するデータベースと，電気自動車と通信して，電気自動車が搭載する電池の種別，エネルギー残量，エネルギー効率，現在位置を受信し，現在位置から前期エネルギー残量およびエネルギー効率によって到達可能な充電スタンドを見つけ，その位置情報と，その充電スタンドが保持する電池のエネルギー残量の情報を電気自動車に送信する充電スタンドロケータとを備える。

※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 57　differential PLSA でのみ抽出された「情報の取得と提供」に関するトピック

（3）differential PLSA の方がトピックの解釈が難しい

以上のように，differential PLSA で抽出されるトピックは，頻度が低い単語でも高い所属確率が割り当てられる傾向にある。そのため，トピックの所属確率が頻度の低い多様な表現に分散しやすく，通常の PLSA で抽出されるトピックよりも意味の解釈が難しい印象がある。実際にこれまで例示したトピックの構成内容を見ても，通常の PLSA と比較して，differential PLSA のトピックは上位語の所属確率が小さく，確率が分散している傾向がある。通常の PLSA は頻度の高い代表的な表現に所属確率が集中するため，解釈はしやすい。しかし，その結果として通常の PLSA で抽出されるトピックは典型的なトピックとなる傾向がある。この問題を解消する手法として differential PLSA を開発した経緯を踏まえると，このジレンマはやむを得ないものである。逆にそれだけ個性的なトピックが抽出されていることを意味するものといえる。もしトピックの構成内容だけでは解釈しにくい場合は，図 56,57 でも示したように，そのトピックに該当する特許の具体的な文章を確認することが効果的である。

9.3.5　differential PLSA のまとめ

differential PLSA は，通常の PLSA で抽出されるトピックが典型的なものになりがちであるという課題に対して，より個性的なトピックを抽出する手法として開発した。通常の PLSA は，データ全体を表現するような再現性の高い代表的なトピックを抽出するため，頻度の高い表現を中心にトピックが構成される傾向があり，結果として典型的なトピックという印象を持つことがある。これに対して differential PLSA は，期待頻度に対する実測頻度の比率の対数を取った共起行列をインプットにすることで頻度は低いが具体度の高い表現にも高い確率を割り当ててトピックを構成することができる。これによって，より個性の強いエッジの立ったトピックを抽出でき，データ全体では埋もれがちな特徴を発見することができる。

したがって，differential PLSA はビジネスの課題解決の場面において，新たな気づきとなるようなインサイトが得られる手法として期待できるが，あくまでもデータ全体を表現するトピックを把握した前提で用いることが望ましい。データ分析の取り組みでは，木を見て森を見ずとならないように，まずは全体像を把握することが第一であり，全体像の現状を把握した上で個別性のある特徴を探ることが重要となる。そのため，実際には通常の PLSA と differential PLSA を併用した分析が有効であると考える。

10.　Nomolytics，PCSA，differential PLSA の比較

ここまで紹介した Nomolytics，PCSA，differential PLSA という 3 つの手法について，その違いと共通する考え方をまとめる。

10.1　各手法の違い

この 3 つの手法の目的と特徴を比較したものを表 20 に示す。Nomolytics は，テキストデータ全体を表現する代表的なトピックを抽出し，そのトピックをベースに傾向の分析や要因関係の分析をすることで，テキストデータ全体に潜む特徴を俯瞰する手法となる。PCSA と differential PLSA は，Nomolytics の中でもトピックを抽出する機能を拡張させた手法となる。Nomolytics がデータ全体を表現するトピックを抽出する手法であるのに対して，PCSA と differential PLSA はより特徴的で個性的なトピックを優先して抽出する手法となり，より深いインサイトの獲得を狙って開発したものである。PCSA は，特徴を見たいターゲットを定め，そのターゲットの特徴を左右する要因となり得るトピックを抽出する手法となる。differential PLSA は，通常の PLSA が頻度の高い要素を中心とした典型的なトピックを抽出しがちであるのに対して，頻度の大小に依存しないより個性的なトピックを抽出し，データ全体では埋もれがちな特徴を発見する手法となる。どの手法が優れているかということではなく，目的に応じて使い分けることが重要である。たとえば PCSA と differential PLSA は深いインサイトの獲得は期待できるが，トピック抽出のインプットとなる共起行列は加工を施しているため，そこで抽出されるトピックはデータ全体を表現するものではなくなっている。先述したように，データ分析の取り組みでは，まずは全体像を把握することが第一である。データ全体を表現するようなトピックを把握し，全体像を俯瞰したいときには，通常の PLSA を適用することが合理的である。したがって，繰り返しになるが，目的に応じた手法の使い分けが重要ということである。

表 20　Nomolytics，PCSA，differential PLSA の比較

	Nomolytics	PCSA	differential PLSA
目的	テキストデータ全体を表現する代表的なトピックを抽出する。そのトピックをベースに，テキスト情報に潜む傾向や要因関係を可視化し俯瞰する。	特徴を見たいターゲットを定め，そのターゲットに影響を与える要因となり得るトピックを優先的に抽出する。	頻度の大小に依存しない，より個性的なトピックを抽出し，データ全体では埋もれがちな特徴を発見する。
PLSAを適用する共起行列の特徴	「単語×係り受け」の行列構成をとり，全データにおいて，それぞれの単語と係り受けの共起頻度（同時に出現する文章数）をカウントした共起行列を採用する。	全データをターゲットが該当するデータと該当しないデータに分割し，それぞれのデータで同じ構成の「単語×係り受け」共起行列を作成し，その2つの共起行列の差分を取った共起行列を採用する。	全データから集計される実測頻度の共起行列に加え，期待頻度の共起行列を作成し，期待頻度に対する実測頻度の比率の対数を取った「単語×係り受け」共起行列を採用する。

10.2　各手法で共通する分析のコンセプト

一方，この 3 つの手法で共通することは，テキストマイニングで抽出した大量の単語をいくつかのトピックに集約し，その集約されたトピックをベースに傾向を可視化することで，大量のテキストデータに潜む特徴をシンプルに理解しようという分析の考え方である。その分析の考え方のイメージを図 58 に示す。テキストデータを分析する際

に以前から使用されているテキストマイニングという手法は，テキストデータの文書情報に含まれる単語を抽出し，その単語を単位とした可視化によってデータの特徴を分析する。しかし，対象がビッグデータとなると，マイニングによって抽出される単語の数は膨大となるため，その単語をベースとした可視化では結果が複雑になり，解釈が難しくなってしまう。そこで，トピックモデルを応用することで，テキストマイニングで抽出された大量の単語をいくつかのトピックに集約する。これにより，従来の単語ベースではなく，トピックをベースとして，テキストデータに潜む特徴をシンプルに分析することができる。つまり，テキストデータからマイニング（採掘）された単語を，トピックというかたまりにリファイニング（精製）し，さまざまな可視化の形にプロセシング（加工）するというアプローチである。これが本節で紹介した Nomolytics，PCSA，differential PLSA に共通する重要なコンセプトとなる。

図 58　トピックをベースとしたテキストデータの分析

おわりに

本節では，大量のテキストデータを分析するうえで，テキストマイニングにトピックモデルを応用した分析アプローチを紹介した。テキストマイニングは，文書情報に含まれる単語を抽出し，その単語を単位とした可視化によってデータの特徴を把握する手法だが，大量のテキストデータとなると結果が複雑になり，解釈が難しくなってしまう。そこで使われ方の似ている単語を集約できるトピックモデルを適用することで，トピックを単位とした可視化を行うことが有効である。これにより，容易にデータ理解を進めることができ，ビジネスの課題解決においてインサイトの獲得もしやすくなる。

一般的なトピックモデルでは，そのインプットとする共起行列のデータは「文書×単語」という Bag-of-Words の形式をとるが，本節では，トピックモデルの PLSA の適用において，「単語×係り受け」という構成の共起行列を採用することを特徴とした，① Nomolytics，② PCSA，③ differential PLSA という 3 つの開発手法を紹介した。それぞれの手法は目的が異なっている。Nomolytics はデータ全体を表現するトピックを抽出して全体像を理解する手法となる。そのトピックをベースに様々な軸で傾向を可視化したり，そのトピックの周辺に潜む要因関係をベイジアンネットワークで分析する。PCSA と differential PLSA はより特徴的で個性的なトピックを抽出することでインサイトの獲得を狙った手法となる。PCSA は特徴を見たいターゲットの要因となるトピックを抽出し，differential PLSA は頻度の大小に依存することのない個性的なトピックを抽出する。また，これらの手法で共通する本質的な考え方は，テキストマイニングで抽出した大量の単語を，いくつかのトピックに集約し，そのトピックをベースに傾向を可視化する

ことで，大量のテキストデータに潜む特徴をシンプルに理解するということである。

一方，ビジネスの課題解決ではデータの理解だけでは不十分であり，そこから有効なアクションが具体的に講じられなければ意味がない。大量のテキストデータをトピックでモデル化し，そのトピックをベースに可視化した結果をいくら提示しても，その先にあるのは「で，だから何？」という反応である。これはデータ分析全般で頻繁に発生する問題でもある。簡単な可視化から高度なモデリングまで，さまざまなデータ分析をして大量の結果を並べてみたものの，そこで終了してあるいはそこで達成感を得て，具体的な課題の解決につなげられていない取り組みは多々存在する。筆者はこれを分析の「So what ？ 問題」と呼んでいる。この問題が生じる要因には，いつの間にか課題解決の手段であるはずのデータ分析が目的となっていたということもよくあるが，そもそも分析やモデル化というものが，データを活用した課題解決において何をしているのか認識されていないことも多い。

モデル化とは「抽象化」をする行為であり，この抽象化は観測データという現象に潜む特徴の理解に役立つ。これは具体的な個別事象ばかり眺めていても見えてくるものではなく，現象の全体をメタ視点で抽象化して，そもそもどんな特徴や傾向が普遍的に存在しているのかということを俯瞰して見極める必要がある。モデル化とは，現象の表面的な部分を並べるのではなく，その奥にある本質的な部分を映すためのプロセスとなる。このようにモデル化は課題における特徴を理解することに役立つが，それだけでは課題の解決にはなっていない。課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るため，その理解した課題の普遍的な特徴から再度「具体化」する必要がある。つまり，もともと具体的な個別事象のデータに対して，特徴を理解しやすくするためにモデル化して「抽象化」するが，そこで捉えた課題の特徴から具体的にどのようなアクションで解決に導けるのか，再び「具体化」を施すことが求められる。課題解決のためのデータ分析における「抽象化」と「具体化」のプロセスのイメージを図 59 に示す。

先述した「So what ？ 問題」は，この再度具体化をする前に，抽象化をした分析で終了していることで発生していることが多々ある。実際に抽象化されたモデルは，特徴の分かりやすさを得た反面，抽象度が高い結果であるため，そのモデルの結果をいくら眺めていても具体的な課題解決の施策を検討することは難しい。そこで再度具体化するプロセスでは，その特徴に該当する個別事象のデータの内容を確認してみる。そうするとその普遍的な特徴が具体的にどのような現象として現れているのかイメージすることができる。課題解決の具体的なアクションを講じるためには，論理思考のスキルも求められるが，仮説を立てられる発想力が問われる。「抽象化」から再度「具体化」するプロセスにおいて，再び個別のデータに立ち返り，その内容を再びじっくりと観察することは，具体的なアクションを講じる発想力を刺激する。たとえば，本節で紹介した Nomolytics の一つ目の分析事例の技術⇒用途の関係分析は，技術に対して新たな用途を探索する分析であったが，これは技術トピックと用途トピックのつながりだけ眺めていても何も浮かばない。しかし，そのつながりを示す個別の特許データを確認し，特に想定外の用途で出願されている特許の記述内容を確認することで，具体的な新規用途アイデアを発想することができる。

また，ビジネスコンサルタントの細谷功氏は書籍『具体⇔抽象トレーニング』[25] において，具体化と抽象化を行き来する問題解決の思考法を提案している。問題解決は問題の発見と解決に分けられ，問題の発見で求められることは具体的事象から本質的な課題をメタ視点で抽象化して抽出することであり，問題の解決で求められることはその問題をどうやって解決するのかという具体化であるとしている。この書籍はデータ分析の書籍ではないが，この抽象化というプロセスはまさにデータ分析におけるモデル化そのものである。

したがって，大量のテキストデータを分析してビジネスの課題解決を検討する際には，まずはその複雑で解釈不能なビッグデータをトピックモデルでシンプルにモデル化し，その抽象化されたトピックをベースに特徴を分かりやすく把握することが有効である。そしてその分析で得られた特徴を，元の個別データの中身も確認しながら再度具体化することで，課題解決の具体的なアクションまで発想していく。このように抽象化で終わらない具体性のある分析アプローチを進めることが実務で活きる取り組みとなってくる。そうした考え方と分析アプローチについて，本節の解説が有益な情報を提供できていれば幸いである。

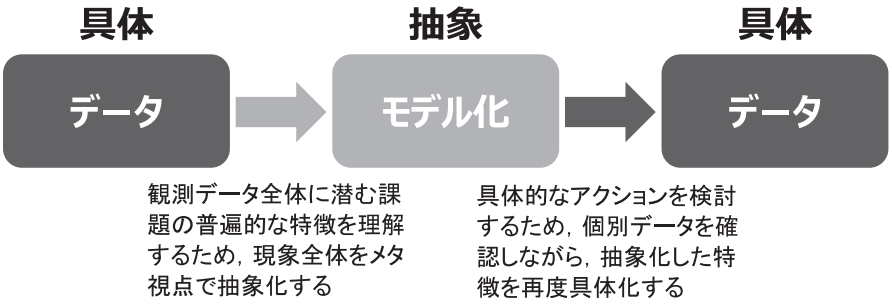


図 59　課題解決のためのデータ分析におけるプロセス

文　　献

1）NITRD (Networking and Information Technology Research and Development) : “The Federal Big Data Research and Development Strategic Plan”, (2016)

2）T. H. Davenport and D. J. Patil : “Data Scientist : The Sexiest Job of the 21st Century, “ Harvard Business Review, Vol. 90, No. 10 (2012)

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）金明哲：『テキストアナリティクス』，共立出版（2018）

5）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)

6）D. D. Lee and H. S. Seung : “Learning the parts of objects by non-negative matrix factorization, “ Nature, vol. 401, pp. 788-791 (1999)

7）T. Hofmann : “Probabilistic Latent Semantic Analysis”, Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, pp. 289-296 (1999)

8)D. M. Blei, A. Y. Ng, and M. I. Jordan : “Latent Dirichlet Allocation,“ Journal of Machine Learning Research, Vol. 3, pp. 993-1022 (2003)

9）柳井啓司：確率トピックモデルによる Web 画像の分類，人工知能学会全国大会論文集，vol. 22（2008）

10）石垣司，竹中毅，本村陽一：百貨店 ID 付き POS データからのカテゴリ別状況依存的変数間関係の自動抽出法，オペレーションズ・リサーチ，Vol. 56，No. 2，pp. 77-83（2011）

11）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）

12）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）

13）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）

14）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）

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－，人工知能学会全国大会論文集，Vol.28（2014）

16）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，人工知能学会全国大会論文集，Vol. 30（2016）

17）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）

18）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）
19）H. Akaike : “Information Theory and an Extension of the Maximum Likelihood Principle”, Proc. of the 2nd International Symposium on Information Theory, pp. 267-281 (1973)
20）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，人工知能学会全国大会論文集，Vol.28（2014）
21）野守耕爾：確率的因果意味解析（PCSA）－テキストデータを用いたターゲット事象の要因トピックの抽出－，人工知能学会全国大会論文集，Vol. 32（2018）
22）特許第 7221526 号：分析方法，分析装置及び分析プログラム（2023 年 2 月 6 日登録）
23）野守耕爾：differential PLSA －テキスト情報の典型的なトピックではないより個性的なトピックの抽出－，人工知能学会全国大会論文集，Vol. 33（2019）
24）特許第 7221527 号：分析方法，分析装置及び分析プログラム（2023 年 2 月 6 日登録）
25）細谷功：『具体⇔抽象トレーニング 思考力が飛躍的にアップする 29 問』，PHP 研究所（2020）