人工知能技術を応用した特許文書分析が生み出す新たな技術戦略の検討

Advanced Technological Strategy Produced by Patent Document Analysis Applying AI Technology

○野守 耕爾 （株式会社アナリティクスデザインラボ）

要旨:テキストマイニングに PLSA（確率的潜在意味解析）とベイジアンネットワークという 2 つの人工知能技術を応用した新たな特許文書分析アプローチを適用事例とともに紹介する. 風や空気に関連する約 3 万件の特許公報の要約文を分析対象とし, PLSA の実行によりその要約内容を数十個のトピックに集約して全体像をシンプルに理解可能にした. またそのトピックを軸として, 技術のトレンドを把握したり, 各企業の開発動向を可視化して競合分析することで提携戦略や競争戦略を検討するヒントが得られた. さらにベイジアンネットワークにより用途と技術の確率的因果関係をモデル化することで, 企業が保有する技術の新たな用途展開を検討するアイデアを発想した.

Abstract: This presentation introduces an approach of patent document analysis and an example applying the approach. The approach is constructed of text mining and two AI technology: PLSA and Bayesian Network. This study analyzes about 30,000 patent document data related to wind and air. Applying PLSA can reduce patent documents to dozens of topics, which is useful to understand the overviews simply. Data aggregation and visualization based on the topics enables to understand technical trends and analyze competitor's technical position leading to the strategy of alliance and competition. Modeling the relationships between usage and technology by Bayesian Network gives awareness of new usage of technology owned by companies.

1. はじめに

企業の技術戦略を検討するうえで, その技術領域の動向を把握し, 自社の技術と他社の技術の特徴を俯瞰して理解することは重要である. 通常, 他社の技術開発動向は機密性が高いため外部から確認することは難しいが, 特許情報はそれを探ることのできる貴重な公開情報である. 特許情報を分析することは企業の技術戦略の検討において有用性が高いことは明らかである.

特許情報分析というと従来パテントマップ [1]と呼ばれるものが代表的であり, これは主に出願人や出願年, 特許分類（IPC, FI, F ターム等）を軸にして特許件数を集計・可視化する. 近年では特許情報の要約文や請求項, 明細書などの文書を分析対象とし, テキストマイニング技術を応用することで, 人間では読み切れない膨大な特許文書の全体像を把握するアプローチもよく採用されている [2]. 現在では複数の企業からこうした分析ツールが販売されており, 分析事例が報告されている [3].

一方, 昨今は第三次人工知能ブームと呼ばれ, 特に注目されている技術はディープラーニングといえるが [4], 他にも人工知能の分野で発展してきた技術には有用なものがあり, テキストマイニングによる単語の抽出と集計に留まっていた特許文書分析も, こうした人工知能技術を応用した新たな分析の検討が進められている [5][6].

本稿では, テキストマイニングに PLSA（確率的潜在意味解析）とベイジアンネットワークという 2 つの人工知能技術を応用することで, 企業の技術戦略の検討において新たな知見の創出が期待できる特許文書分析のアプローチとその分析事例を紹介する.

2. 従来の特許文書分析

特許文書にテキストマイニングを応用した従来の分析において, その分析目的によく取り上げられるものとしては, ①全体像を把握する, ②トレンドを把握する, ③競合他社の動向を把握する, ④用途と技術の関係を把握するといったものが挙げられる. アウトプットとしてよく用いられるものを図 1 に示す.

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

②トレンドの把握は, 今後成長が見込まれる技術や, 逆に衰退している技術を把握し, 研究開発戦略を検討していくための分析である. 抽出した単語の出現頻度を出願年で集計することもあるが, 単語ベースでは複雑になりすぎるため, しばしば図 1(C)のように単語や係り受けを人がグルーピングして意味性のあるカテゴリを形成し, 図 1(D)のようにカテゴリ別に該当する特許件数を出願年で集計してそのトレンドを把握する.

③競合他社の動向把握は, 他社と差別化する研究開発戦略を検討したり, 自社技術のライセンス先候補や提携候補となる企業を検討するうえでニーズがある分析である. 図 1(E)のように単語と出願人を同じ平面上にマッピングし, 出願人と近くに位置する単語から各出願人の技術開発動向を把握する.

④用途と技術の関係把握は, 特に自社技術の新たな用途展開を探索するうえでニーズがある分析である. 特許の要約文で記述されていることの多い「課題」と「解決手段」という 2 つの項目に着目し, それぞれの文書をテキストマイニングして図 1(C)のようにカテゴリを形成し, 図 1(F)のように課題のカテゴリと解決手段のカテゴリに該当する特許件数をクロス集計することにより, 用途と技術の対応関係を把握する.

これらの分析は, 人間ではなかなか読み切れない特許文書の全体像を把握するうえで有効な手段であるが, 以下のような課題もあるといえる.

(1) 基本的に単語をベースにした分析であるため, 結果が複雑で考察しにくい
(2) カテゴリの設定が主観的で作業負荷も大きい
(3) 用途と技術の関係は単純なクロス集計で統計的な関係を分析できていない

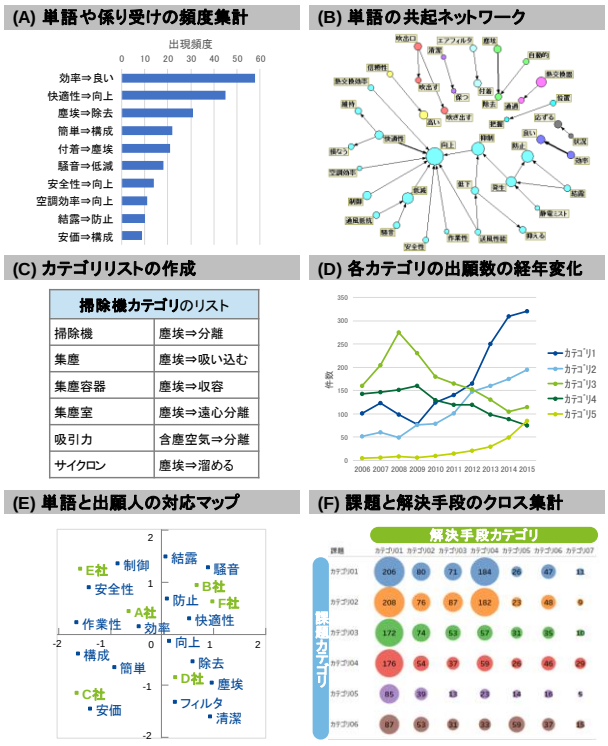

図 1　従来の特許文書分析におけるアウトプット例

3. 人工知能技術を応用した新たな文書分析技術

本稿ではこうした従来の特許文書分析の課題に対して人工知能技術を応用することで解決を検討する．上記の(1)(2)の課題については単語をクラスタリングする人工知能技術として PLSA（確率的潜在意味解析）を，(3)の課題については要因関係をモデリングする人工知能技術としてベイジアンネットワークを適用する．

PLSA(Probabilistic Latent Semantic Analysis)[7]は，共起行列と呼ばれる行列データをインプットとし，行の要素と列の要素の背後にある共通特徴となる潜在クラスを抽出する手法であり，トピックモデルと呼ばれる技術の一つで，クラスタリング手法としても適用される．

ベイジアンネットワーク[8]は，複数の変数の確率的な因果関係をネットワーク構造で表わし，ある変数の状態を条件として与えたときの他の変数の起こりうる確率（条件付確率）を推論することができる確率モデルである．

図 2 に示すように，従来のテキストマイニング技術に加え，クラスタリング技術の PLSA とモデリング技術のベイジアンネットワークを連携させた文書分析技術として，Nomolytics (Narrative Orchestration Modeling Analytics) を開発した（特許第 6085888 号，商標出願中）．本技術では，まずテキストマイニングにより文書から単語を抽出し，単語間の共起頻度をデータ化した共起行列を作成する．次にその共起行列に PLSA を適用し，使われ方の似ている単語をトピックにまとめ上げ，全文書データに対して各トピックの該当度も計算する．最後にベイジアンネットワークによって，トピック間あるいは他の属性情報との間の確率的な因果関係をモデル化する．

こうした 3 つの技術を組み合わせることで，膨大な文書データをいくつかのトピックという人間が理解しやすい形に整理でき，ベイジアンネットワークによってその文書データに潜む複雑な要因関係を構造化できる．本技術は文書データであればあらゆる分野で適用でき，例えば旅行の口コミデータに適用して地域観光のマーケティングを検討する事例がある[9]．本稿ではこれを特許文書に適用した事例について紹介する．

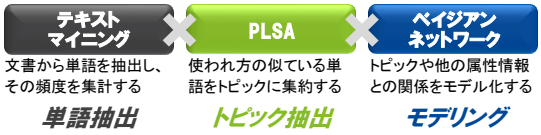

図 2　新たな文書分析技術：Nomolytics

4. Nomolytics を応用した特許文書の分析方法

Nomolytics を特許文書に適用した分析は，(A)用途と技術のトピック抽出，(B)トピックの特徴集計，(C)用途と技術の関係分析という 3 つのステップから成る．ステップ(A)では，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 を適用して，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．ステップ(B)では，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）を計算し，そのスコアを出願年で集計してトレンドを分析したり，そのスコアを出願人で集計して，各企業の技術開発動向を可視化する．ステップ(C)では用途トピックと技術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，ある技術の新しい用途展開を探索する．

5. Nomolytics を応用した特許文書の分析事例

本稿の紹介事例では，要約と請求項に「風」「空気」を含む 10 年分（出願日が 2006 年 1 月 1 日～2015 年 12 月 31 日）の特許公報 30,039 件を分析対象とした．

5.1. 用途と技術のトピック抽出

特許の要約文から「課題」に関する項目で記述された文章と「解決手段」に関する項目で記述された文章を抽出し，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を抽出した．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．

続いて PLSA のインプットとする共起行列を作成した．本来の PLSA では，文書（行）×単語（列）という構成の共起行列を用いるが，Nomolytics では，単語（行）×係り受け（列）という構成でそれぞれの共起頻度を集計した共起行列を用いる．PLSA は行と列の背後にある共通特徴を抽出するが，こうした構成の共起行列に適用することで，単語という話題の観点となる軸と，係り受けというその観点の具体的な内容となる軸でその共通特徴を抽出することになり，より文脈上近しい単語と係り受けでまとめられた解釈のしやすいトピックを抽出することができる．なお本事例では頻度 10 件以上を対象とし，「課題」の文章からは単語（3,256 語）×係り受け（2,084 表現）の共起行列を，「解決手段」の文章からは単語（5,187 語）×係り受け（7,174 表現）の共起行列を作成した．

この共起行列に PLSA を適用し，「課題」からは用途のトピックを，「解決手段」からは技術のトピックを抽出した．なお PLSA は予めトピック数を設定する必要があり，また初期値により解が異なる特性がある．そこでトピック数を 1 刻みで変化させ，それぞれのトピック数に対して PLSA を初期値を変えて 5 回ずつ実行し，それぞれの解を情報量基準 AIC で評価して最も評価の良い解を採用した．その結果，用途については 25 個のトピックが，技術については 47 個のトピックが得られた．なお PLSA の実行には 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[10]という PLSA を拡張さ

せた同様の分析機能を使用した．

PLSA のアウトプットは，①各トピックにおける行要素（単語）の所属確率，②各トピックにおける列要素（係り受け）の所属確率，③各トピックの存在確率，という 3 つの確率が計算される．抽出された用途と技術のトピックの内容の例を表 1 に示す．単語と係り受けは所属確率の高い順に並べている．用途トピック（表 1 左）では，単語は，加湿装置，水，供給，加湿，カビなどが，係り受けは，加湿装置の提供，加湿器の提供，ミスト発生装置の提供，水の供給，細菌の繁殖などが関係しているので，この結果は加湿に関するトピックであると解釈できる．技術トピック（表 1 右）では，単語は，送風機，塵埃，掃除機，分離，吸い込む，集塵部などが，係り受けは，塵埃の分離，分離する塵埃，塵埃を含む，吸い込む塵埃，含む空気，空気の分離などが関係しているので，この結果は塵埃の分離に関するトピックであると解釈できる．このように解釈をつけた 25 個の用途トピックと 47 個の技術トピックの一覧をそれぞれ表 2,3 に示す．

表 1　用途トピックの例

用途トピックU04				技術トピックT32			
確率	単語	確率	係り受け	確率	単語	確率	係り受け
5.5%	加湿装置	6.8%	加湿装置-提供	5.5%	送風機	2.1%	塵埃-分離
3.7%	水	3.1%	加湿器-提供	5.2%	塵埃	1.7%	分離-塵埃
3.3%	供給	2.9%	ミスト発生装置-提供	4.1%	掃除機	1.7%	塵埃-含む
2.4%	加湿	1.9%	水-供給	3.6%	分離	1.5%	吸い込む-塵埃
2.3%	カビ	1.7%	細菌-繁殖	3.5%	吸い込む	1.3%	含む-空気
2.1%	加湿器	1.5%	加湿-行う	2.3%	集塵部	1.0%	空気-分離
…	…	…	…	…	…	…	…

表 2　用途トピックの一覧

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

表 3　技術トピックの一覧

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

5.2. トピックのスコア計算

続いて全特許データに対して，抽出された各トピックのスコア（該当度）を計算する．1 件の特許データには複数の文章で構成されているため，まず文章単位（句点で区切られた文章単位）に各トピックのスコアを計算し，それを特許単位に集約する．文章 S_h におけるトピック T_k のスコアは $P(S_h|T_k)/P(S_h)$ で定義する．これはトピックを条件とすることで文章の発生確率が何倍になるのかを示し，そのトピックをよく話題にしている文章ほど高くなる．以下，$P(S_h|T_k)$ と $P(S_h)$ の計算について説明する．

$P(S_h|T_k)$ については，文章 S_h を単語で定義される文章 Sw_h と係り受けで定義される文章 Se_h に分解し，それぞれについて $P(Sw_h|T_k)$ と $P(Se_h|T_k)$ を計算し，それらを一つに統合して $P(S_h|T_k)$ を計算する．$P(Sw_h|T_k)$ と $P(Se_h|T_k)$ は式(1)(2)で計算される．単語 W_i と係り受け E_j が含まれる文章の数をそれぞれ $n(W_i)$ と $n(E_j)$ とすると，$P(Sw_h|W_i)$ は $n(W_i)$ の逆数，$P(Se_h|E_j)$ は $n(E_j)$ の逆数として計算される．$P(W_i|T_k)$ と $P(E_j|T_k)$ はそれぞれ PLSA の実行結果によって得られている．$P(S_h|T_k)$ は式(3)で計算され，$P(S_h|Sw_h)$ と $P(S_h|Se_h)$ は文章 S_h において重みは同じであるためそれぞれ 0.5 とする．$P(S_h)$ は式(4)で計算され，$P(T_k)$ は PLSA の実行によって得られている．

$$P(Sw_h|T_k) = \sum_i P(Sw_h|W_i)P(W_i|T_k) \quad (1)$$

$$P(Se_h|T_k) = \sum_j P(Se_h|E_j)P(E_j|T_k) \quad (2)$$

$$P(S_h|T_k) = P(S_h|Sw_h)P(Sw_h|T_k) + P(S_h|Se_h)P(Se_h|T_k) \quad (3)$$

$$P(S_h) = \sum_k P(S_h|T_k)P(T_k) \quad (4)$$

以上から $P(S_h|T_k)/P(S_h)$ で定義されるスコアを文章単位に計算し，それを特許単位に見たとき，各トピックのスコアの最大値をその特許のトピックスコアとして採用した．さらにこのスコアの閾値として 3 を設定し，各特許データに対してそのトピックの該当有無を示す 0,1 のフラグ情報を付与した．$P(S_h|T_k)/P(S_h)$ で定義したスコアは 1 が基準となるが，本事例では各トピックの特徴をより濃く抽出するため，このスコアの分布や実際の文章内容を確認しながら妥当性も検討し，基準の 3 倍と厳しく設定した．

5.3. トピックのトレンド分析

特許データの出願年の情報と，各トピックのフラグ情報から，トピックのトレンドを分析した．具体的には出願年 Y とトピック T の関連度を示す指標として $P(Y|T=1)/P(Y)$ を計算し，この値の経年変化を可視化した．技術トピックにおいて，2013 年からの上昇率が高い上位 5 つのトピックのトレンドを図 3 に示す．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．

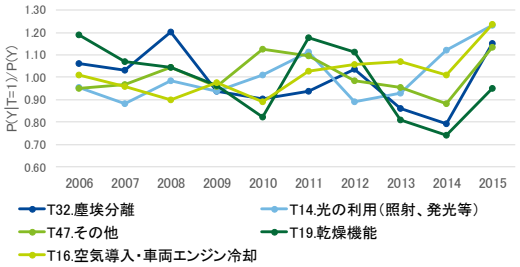

図 3　2013 年から上昇率トップ 5 の技術トピックトレンド

5.4. トピックの競合分析

特許データの出願人の情報と，各トピックのフラグ情報から，各トピックにおいて出願人の動向を可視化する競合分析を行った．具体的には出願人 X とトピック T の関連度を示す「シェア」と「注力度」という 2 つの指標を計算して分析した．シェアとは，P(X|T=1)で定義され，そのト

ピックが該当する特許の中でのその出願人の出願割合を示す．注力度とは，P(T=1|X)で定義され，その出願人が出願した特許の中でのそのトピックの該当割合を示す．本事例では縦軸にシェア，横軸に注力度を設定し，トピックごとに各出願人のポジショニングを可視化した．トレンド分析で短期的に上昇していた技術トピック T32.塵埃分離を例とした結果を図 4 に示す．

図 4 より，C 社は高水準のシェアを獲得しつつ，注力度は他社と比べてとても高く，高い技術力を保有している可能性がある．今後はよりシェアを伸ばすことで高シェア高注力度のポジションを確立できる．一方 A 社と B 社もシェアは高いが，C 社に注力度で劣る．例えば規模は中程度だが比較的注力度が高く高い技術力があると思われる E 社，G 社，I 社などと連携することで，C 社の上のポジションを狙うことができる可能性もある．このように，塵埃分離に関する技術は，1社の注力度が高いものの，他にもある程度のシェア・注力度を保有する企業が何社か存在し，またトレンドも近年ホットであるため，今後企業連携などの動きも十分考えられる領域と推察できる．

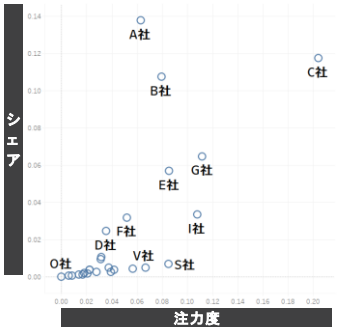

図 4 技術トピック T.32 塵埃分離における出願人動向

5.5. 用途と技術の関係分析

用途トピックと技術トピックのフラグデータを用いて，技術トピックに対する用途トピックの関係構造をベイジアンネットワークでモデル化した．構築したモデルを図 5 に示す．なおベイジアンネットワークの適用には BayoLink（株式会社 NTT データ数理システム）を使用した．

図 5 のモデルを用いることで，ある技術トピックに対して確率統計的に関係の強い用途トピックを把握できる．例えば技術トピック T18.除湿では，用途トピック U05.乾燥機能（衣類など）や U12.制御（冷媒回路等）に対して関係が見られた．この技術と用途の関係性の分析結果を用いて，技術の新しい用途を具体的に検討していく．

例えば，技術トピック T18.除湿は，用途トピック U05.乾燥機能と関係が強い結果が得られたが，ある出願人 X に着目すると，X 社の T18.除湿に該当する特許のうち，U05.乾燥機能に該当するものはほとんど存在しなかった．つまり全体での関係性を見れば，この X 社の保有している T18.除湿に関する技術はもっと U05.乾燥機能の用途に展開できる可能性があるといえる．

さらに実際の特許文書を確認することで，この新規用途探索の分析をより深めていく．まず T18.除湿の技術が U05.乾燥機能の用途を想定して出願されている特許の代表例としてドラム式洗濯乾燥機の特許があり，特許文書を確認すると，例えば洗濯物を短い時間でムラ無く乾燥させ，乾燥工程の時間を短くするための除湿技術が求められている．一方出願人 X が出願している T18.除湿の技術に関する特許には，インクジェットプリンタに関する特許がある．インクジェットプリンタでは，吹き付けるインク液にムラが出ないようにすること，またそのインク液を吸収した紙が湿度のムラによって波打たないように乾燥処理することが求められている．X 社はプリンタの中で，紙に残った余分なインク液を加熱して蒸発させて，その蒸発による湿気を吸引ファンで取り除くことで，インク液が不均一にならないように乾燥処理をする技術を特許として出願している．X 社は洗濯乾燥機の製造はしていないが，プリンタという空間の中で，インク液を吸収した用紙の湿気をムラなく取り除いて紙の波打ちを防ぐ乾燥処理技術は，例えば洗濯乾燥機の中で洗濯物をムラなく効率的に乾燥させることに応用できる可能性もある．これはあくまで分析結果から発想したアイデアであり，現実性は検討していないが，こうした分析により新しい用途展開の気づきが得られることが期待できる．

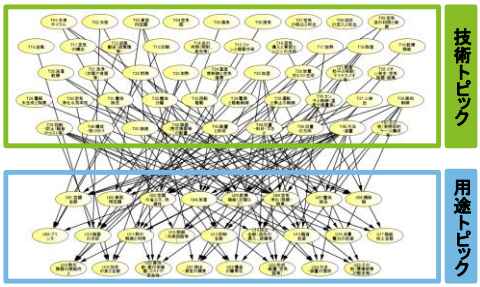

図 5 技術に対する用途の関係モデル

6. まとめ

本稿では，テキストマイニングに PLSA とベイジアンネットワークという 2 つの人工知能技術を応用した新たな文書分析技術（Nomolytics）と，それを特許文書データに適用した分析事例を紹介した．Nomolytics を適用した特許文書分析のメリットには，①単語ではなくトピックをベースにした分析を実行することで，膨大な特許文書に潜む特徴を分かりやすく理解することができること，②技術と用途の統計的な関係を把握して，技術の新規用途のアイデアを発想できることが挙げられる．こうした分析結果を活用することで，企業の技術戦略の検討において新たな知見の創出が期待できる．

参考文献

1) 新井喜美雄（2003）「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.
2) 安藤俊幸（2009）「テキストマイニングと統計解析言語 R による特許情報の可視化」，情報管理，Vol.52，No.1，pp.20-31.
3) 山中なお（2010）「知的財産戦略に資する特許情報分析事例集」，特技懇，No.259，pp.82-84.
4) 松尾豊（2015）「人工知能は人間を超えるか ディープラーニングの先にあるもの」，角川 EPUB 選書.
5) 安藤俊幸（2016）「機械学習を用いた効率的な特許調査方法」，Japio YEAR BOOK 2016，pp.150-161.
6) 岩本圭介（2016）「特許文献から技術動向を把握するためのマイニング手法」，Japio YEAR BOOK 2016，pp.198-203.
7) Hofmann, T. (1999) “Probabilistic latent semantic analysis,” Proc. of Uncertainty in Artificial Intelligence，pp.289-296.
8) 繁桝算男・植野真臣・本村陽一（2006）「ベイジアンネットワーク概説」，培風館.
9) 野守耕爾・神津友武（2016）「口コミビッグデータに人工知能を応用した地域観光の次世代マーケティング」，2016 年度人工知能学会全国大会論文集.
10) 若杉徹・高橋勲男（2014）「医薬品調剤履歴に関する確率的構造解析に基づく適応症の推定」，2014 年度人工知能学会全国大会論文集.