Original Paper

J Jpn Ind Manage Assoc 68, 161–170, 2017

Knowledge Extraction from Textual Patent Data Using a Random Forest

Takumi Tsumura, Fumiaki Saitoh and Shohei Ishizu [†1]

Abstract

In the development of the new product and the management of technologies, it is important to make decisions based on the information about technological trends, such as patents. The relationship between the application range and field of technology has become complicated due to the use of increasingly sophisticated technologies in recent years. Along with this, knowledge sharing between decision-makers and engineers has become increasingly difficult. Under these circumstances, there is a need for an efficient tool capable of supporting knowledge sharing among engineers and managers whose expertise are in different areas. The purpose of this study is the knowledge extraction of patent data through constructing a text-mining patent map using the random forest methodology. By extracting the knowledge from the results learned using the technological information obtained through random forest as the supervised signal, we have constructed patent maps that take the relationships between technical fields into consideration. The method we propose creates a patent map by combining two-dimensional kernel density estimation and multi-dimensional scaling based on a similarity matrix calculated applying the internal components of the random forest. We also deal with calculation results of non-negative matrix factorization as a random forest input, thereby avoiding the vulnerability of the random forest noise variables. Non-negative matrix factorization is useful for interpreting the important variables extracted in the random forest. In the experiment, we confirmed the behavior of the method proposed using Japanese patent data.

Key words: patent-map, random forest, non-negative matrix factorization, management of technology, text mining

[†1] Aoyama Gakuin University
Received: August 17, 2016
Accepted: June 15, 2017

原 著 論 文

ランダムフォレストを用いた特許に関する文書データからの技術適用領域に関する知識抽出

津村 拓海，齊藤 史哲，石津 昌平[†1]

技術情報の複雑化・多様化に伴い，専門知識が異なる技術者や意思決定者の間で技術傾向に関する情報の共有が難しくなっている．こうした中で近年，テキストマイニングに基づいたパテントマップが注目されており，これを用いることで特許データから技術に関する知識の抽出・可視化が実現されている．本研究は，複数の分野を跨いで実用の可能性に富む技術に関する情報の可視化およびそれらに関する知識抽出を行うものである．提案法では出現した単語の情報を属性とし，技術分野などに関する情報をクラスとしたランダムフォレストの学習結果の内部モデルに対する解析を通じて知識抽出を行う．知識抽出にあたり，ランダムフォレストの内部モデルから特許文書間の類似度行列を作成し，多次元尺度構成法によるマッピングを行っている．ここで得られた出力結果に対してカーネル密度推定を施すことによって，適用領域を考慮した文書ベースのパテントマップが構築できる．また，非負値行列因子分解により，単語の類似性に基づいて次元縮約することでランダムフォレストの内部モデルの解釈を与えた．分析対象には Web 上に公開されている特許に関する文書データを用いて分析を行った．

キーワード： ランダムフォレスト，パテントマップ，非負値行列因子分解，技術経営，テキストマイニング

1 はじめに

近年における科学技術の急激な発展に伴い，技術に関する情報は複雑化・多様化している．このため，技術戦略の策定において要求される専門知識は多岐にわたる．こういった状況下では，経営上の意思決定者と技術者の関係や専門分野が異なる技術者間の関係などといった専門知識やバックボーンが大きく異なるにも関わらず，協働して共通の目標を達成しなければならない状況がしばしば発生する．経営に関する意思決定においては，競合他社との差別化や参入障壁の検討といった現状の技術動向に関する分析が重要になる．特に技術経営の観点から考えると，特定技術に限定した狭い業務領域における問題解決よりも，技術と経営が入り混じった広い業務にまたがる問題解決へのニーズが強いと考えられており，組織内における技術に関する知識共有は今後ますます重要視されるといえる．

こうした中で近年，技術動向やその傾向に関する理解や認識の共有を目的として，パテントマップなどの分析ツールが重要視されている．パテントマップとは，特許に関連する様々な情報に対して 2 次元マッピングによる可視化を通じて技術に関する知識を抽出するツールである．パテントマップは出願者数や件数といった技術動向に関する統計データに基づくもの（統計解析型）と出願者間の関係などの非数値の情報を対象としたもの（非統計型），文書データに基づくもの（テキストマイニング型）に大別される [1]．本研究は，技術内容に関する文書データに基づいて技術動向や分野間の差異に関する知識の抽出を目指すものであり，テキストマイニング型のパテントマップに関する研究である．

特許に関する情報をマイニングの技術によって分析する試みは近年盛んにおこなわれており，これらの文書データを対象とした自然言語処理・テキストマイニングを適用する試みがなされている [2]～[7]．パテントマップにおいても，例外ではなく，同様な試みがなされている．テキストマイニング型に限らず，いずれのパテントマップも，技術経営における意思決定を支援するためのツールとして盛んに利用されており，これらは今後さらに重要視されと考えられる．多くの場合，特許に関する文書データに基づいたパテントマップには，主成分分析や多次元尺度構成法，自己組織化マップなどの統計的データ分析ツールが広く用いられており，予測対象を持たない（教師なし）データ分析の枠組みで議論される [8]．

その一方で，各技術（特許）を何らかの分類基準（教師情報）を設けてモデル作成ができれば，その基準を考慮した特許の理解や解釈が可能になる．たとえば，技術分野の情報を教師信号として扱えば，技術分野に基づいたモデルにより他分野に対する応用の可能性に関する議論が可能になる．また，年代の情報に基づいたマッピングができれば，新しい技術であるにも関わらず，以前から先取りされていた可能性がある技術や，その逆の可能性がある技術を視認することが可能にな

[†1] 青山学院大学
受付：2016 年 8 月 17 日，再受付（2 回）
受理：2017 年 6 月 15 日

る．このように，分類対象であるクラスの情報を考慮したモデルを用いてマップが作成できれば，単に文書間の類似度に基づいて分析されてきたこれまでのパテントマップよりも，技術経営における意思決定を支援する上で有益な知識の獲得が期待できる．

特許データに教師ありモデルを適用した研究に目を向けると，特許の質に関する指標を出力（予測対象）としたデータマイニングツールを適用した研究が行われている [9],[10]．これらの研究の多くは，ニューラルネットワーク (Artificial Neural Network, ANN) やサポートベクターマシン (Support Vector Machine, SVM) に代表される教師あり学習モデルを利用している．一般に，ANN や SVM はブラックボックスとして扱われるモデルであり，予測や分類に至る規則性や法則性の理解は難しく，データマイニングの分野ではこのようなモデルの学習結果から知識抽出を試みる研究がされてきた．

本研究では教師あり学習モデルとして ANN や SVM と同等以上の識別性能を持ち，近年多くの分野で高い成果を収めているランダムフォレスト (Random Forest, RF) に注目する．RF は複数の決定木から構成されたモデルであるため，これらの反応を調査することによって単なる予測モデルとしてではなく，特許に関する知識抽出が可能となる．以下本論文では，RF の特許の文書データへの適用を通じて，技術戦略における意思決定支援ツールを提案する．

2 基礎事項の準備

ここでは，準備として本論文中で扱う分析手法について概説する．

2.1 ランダムフォレスト

2.1.1 決 定 木

まず，本研究で利用するランダムフォレストの基礎をなす決定木にについて概説する．決定木とは学習に基づくデータマイニングツールの一つであり，主にデータの分類・予測などに利用される．決定木は図 1 に示す通り，複数の条件分岐を段階的に重ねた閉路を含まない有向グラフ（木）を，データに当てはめることによって分類・識別を実現する．これはノードとリンクから成り，頂上の 1 つのノードは「根ノード」，終端の四角いノードは「葉ノード」と呼ばれている．

頂上の根ノードが最初の条件分岐を表しており，終端の葉ノードがデータが分類された結果を表している．様々な学習アルゴリズムが提案されており，ランダムフォレストでは最も代表的なアルゴリズムの 1 つである CART を改良した学習による決定木が用いられる．

2.1.2 ランダムフォレストの概要

ランダムフォレスト (Random Forest, RF)[11] は複数の決定木によって構成される学習モデルであり，ブートストラップ型のアンサンブル学習に分類される．アンサンブル学習とは，複数の弱学習器を個別に学習させ，それらの予測結果を平均や多数決などに基づいてモデル全体の出力とすることによって個々の学習器よりも高精度なモデルを構築する方法である．構成要素である弱学習器は識別性能が決して高いとはいえないが，モデル全体としては汎化性能を維持しつつ，高い識別能力を持つことが，アンサンブル学習の特徴である．

特に RF は高い予測・識別精度を持ち，高次元データに対しても頑健であり，過学習などのリスクが低いことなど様々な利点を有している．このため，近年様々なデータ解析において応用されており，広く研究されている．RF では，多様性を持たせた複数の決定木を構築するにあたって，データセットからブートストラップ [12] と呼ばれるリサンプリングを用いてデータを複製する．それぞれの決定木に対してこの方法を適用し，複製されたデータに対してランダムに変数選択を行った決定木を学習させる．

ここでは，学習に用いる N 個のデータを $(\boldsymbol{x}_i, y_i), i \in \{1, 2, \ldots, N\}$ とする．$\boldsymbol{x}_i \in \mathbb{R}^d$ は d 次元の説明変数（属性）であり，y_i はクラスである．本研究でいうと，前者が特許に関する文書データで後者が予測対象である技術分野などに対応している．RF を構成する B 個の木を $T_b, b \in \{1, 2, \ldots, B\}$，それに対するそれぞれの識別および予測の結果を $h(\boldsymbol{x}_i; T_b)$ とすると，RF の出力は分類のタスクにおいてはこれらの多数決による識別結果を，回帰のタスクにおいてはこれらの平均値を返す．

上記の表現に基づくと，RF のアルゴリズムは以下のように表現できる．

.....

Step.1 $b = 1$ から B まで以下の処理を繰り返す．

(a) N 個のデータからブートストラップサンプル Z_{boot} を求める．

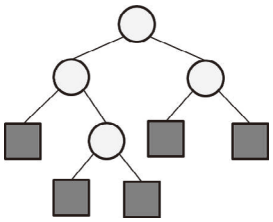

図 1 決定木の概略図

(b) d 次元のデータから $k(<d)$ 個の変数をランダムに選択し（**注 1**），(**a**) で求めた Z_{boot} を用いて木 T_b を学習させる

Step.2 **Step.1** で学習した結果を用いて，それぞれの木による識別・予測の結果 $h(\boldsymbol{x};T_b)$ を出力する．

Step.3 B 個の木 T_b のそれぞれの識別結果 $h(\boldsymbol{x}_i;T_b)$ に基づいて，各データ $\boldsymbol{x}_i$ に対し，回帰では (1) 式，分類では (2) 式を用いて予測結果 $\hat{f}(\boldsymbol{x}_i)$ を求める．

$$\hat{f}(\boldsymbol{x}_i)=\sum_{b=1}^{B}h(\boldsymbol{x}_i;T_b)/B \tag{1}$$

$$\hat{f}(\boldsymbol{x}_i)=\arg\max_{1\le j\le K}|C_j| \tag{2}$$

ここでは，K 種類のクラスを持つデータにおいて，$|C_j|$ はクラス $C_j, j\in\{1,...,K\}$ を返した木の数を表している．

なお，RF の概要は図 2 に示すとおりである．ここでは，学習データセットから B 個の Sub Set を構築し，それぞれの Sub Set に対して決定木を学習させることを表している．この結果より獲得された B 個の予測結果の多数決によって最終的な予測結果が出力される．図中における各 Sub Set を抽出する際には，先述のブートストラップサンプリングによってインスタンスを選択し，さらに素性（変数）をランダムに選択している．

2.2 非負値行列因子分解による次元縮約

RF に対して単語の頻度行列などの，大規模疎行列を適用すると，木を学習させる際に学習対象としてサンプリングされたデータ中に対象の単語が含まれないケースがしばしば発生する．このようなデータに対しては，木の学習が困難であるため RF の有効活用は難しい．そこで，あらかじめ何らかの方法で，次元縮約などの変数変換を施す必要がある．

本研究では，文書データにおける単語の頻度行列などの大規模疎行列の次元縮約において広く利用されている，非負値行列因子分解（Non-negative Matrix Factorization, NMF）[13] を利用する．NMF とは，各インスタンスに対する単語の出現頻度を表す行列を，基底に基づいた行列の積として近似的に表現する次元縮約およびクラスタリングの手法である．すなわち，NMF ではインスタンスと基底からなる行列と基底と単語からなる行列の積として近似的に分解する．

図 3 に示すとおり，基底数を K，データの次元数を I，データサンプル数 J とした際に，データに対応する $I\times J$ 行列 X を，$I\times K$ 行列 T と $K\times J$ 行列 V を用いて $X\simeq TV$ となるように近似する．文書行列 X の (i,j) 成分 x_{ij} は，縮約された行列 T の第 i 行ベクトルと係数行列 V の第 j 列ベクトルによって，$\hat{x}_{ij}=\boldsymbol{t}_i^T\boldsymbol{v}_j$ として推定される．なお本研究では，基底数 K が縮約後の次元に対応しており，本研究中では，行列 V が RF における特徴量として扱うことにする．

NMF の学習結果では行列の誤差 $\|X-TV\|$ を評価値として，これの局所最小値を得るような行列の変換がなされる．ここでは，変数に対応する行列 T における縮約後の軸で類似した変数の重みを大きくし，なおかつ，インスタンスに対応する行列 V における対応する軸では類似したデータの重みが大きくなるような写像を得る．NMF では，扱う対象データの全要素に対して非負値という制約を設けることで，軸の直交を回避している．このため，軸が直交する潜在意味解析（Latent Semantic Analysis, LSA）との比較において，縮約後の変数に対して解釈を与えやすく意味付けが比較的容易である．

NMF のアルゴリズムの概要は以下の通りである．

Step.1 行列 T および V の各要素を非負値の乱数によって初期化する．

Step.2 予め指定した更新回数 $ite_{\max}$ まで，次式によって t_{ik}, v_{kj} を更新する（**注 2**）．

$$t_{ik}\leftarrow t_{ik}\frac{\sum_j x_{ij}v_{kj}}{\sum_j \hat{x}_{ij}v_{kj}} \tag{3}$$

$$v_{kj}\leftarrow v_{kj}\frac{\sum_i x_{ij}t_{ik}}{\sum_i \hat{x}_{ij}v_{ik}} \tag{4}$$

ここでは，x_{ij}, t_{ik}, v_{kj} はそれぞれ単語の頻度行列 X(左辺の元データ) の要素，右辺左の行列 T の要素，右辺右の行列 V の要素に対応している．i はデータのインデクス，j は単語のインデクス，k は縮約後の因子変数 (トピック) のインデクスを表している．

NMF の特許・技術データへの応用は豊田ら [14] によってなされており，技術に関するトピック抽出において既に効果は確認されている．本研究では，ランダムフォレストの属性として利用することによって，NMF によって抽出されたトピックが技術領域の分析に及ぼす影響を確認する．

3 RF を用いた特許情報からの知識抽出

ここでは，特許データを対象とした際の RF の学習結果から知識を抽出するための方針および方法について述べる．

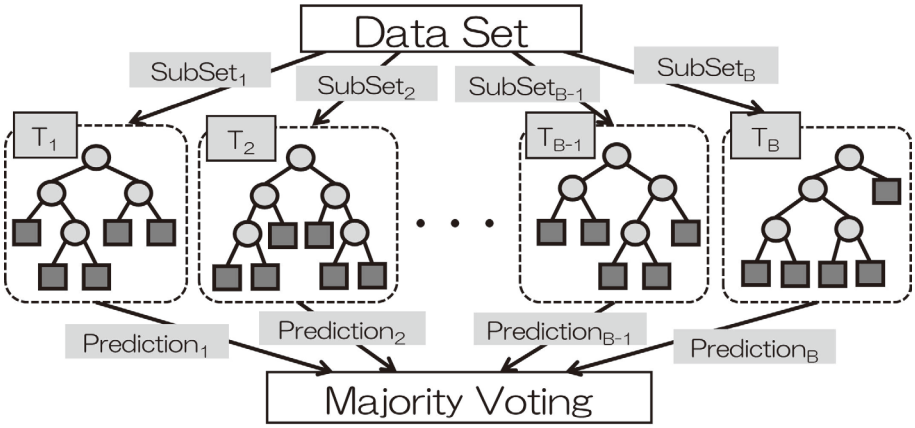

図 2　ランダムフォレストの概略図

3.1　提案法における処理手続き

ニューラルネットや SVM などに代表される教師あり学習モデルの多くは，モデル内部の予測プロセスがユーザにとって把握し難いため，しばしばブラックボックスとして扱われる．データマイニングの分野では，このようなモデルから，知識を抽出する試みが広く行われてきた．本研究においても，これらの試みと同様に教師あり学習モデルである RF の内部モデルから知識抽出を行う．この手続きを通じて，学習した特許に関するデータから技術情報の理解および知識共有の支援を試みる．

提案法における処理の流れは以下の通りである．

提案法における処理手続き

Step.1 特許データに対して形態素解析を実施し，文書データを単語行列に変換する．

Step.2 単語行列を NMF による次元縮約により，低次元密行列に変換する．

Step.3 Step.2 で得られたデータに対して RF を学習させ，技術の分野などをクラスとした分類モデルを構築する．

Step.4 Step.3 にて獲得した RF のモデルから知識抽出を行う．

(a) RF の内部情報を用いたパテントマップを構築する (**3.2**)．

(b) RF の反応に基づいた変数の重要度を算出し解釈を与える (**3.3**)．

特許情報のマイニングにおいてトピック抽出に利用されてきた NMF を，本研究では新たに RF の文書分類における素性とし特許の分類モデルを構築している．さらに，ここで構築したモデルからの知識抽出として，Step.4 における分析手続きを新たに技術の適用領域の理解支援に応用する．

3.2　RF による技術の適用領域を考慮したマッピング

3.2.1　マッピングの指針

各データ間の類似度行列が獲得できれば，多次元尺度構成法 (Multi-dimensional Scaling, MDS) によってデータの低次元空間へのマッピングが可能になる．PCA や SOM と同様に，文書に対するパテントマップでは MDS を適用されることが多い．ここでは，コサイン類似度などを用いて文書間の類似度行列を構築し，それに対して MDS が適用される．コサイン類似度を用いた MDS によるパテントマップでは，低次元マッピングに反映される情報がデータ間の類似度のみであり，クラスの情報が反映されていない．非線形性の強いデータではマッピングの結果は特徴量が類似していても領域によって識別対象が異なるケースがあり，適切な可視化を実現できず解釈が難しくなるケースがある．

RF では，学習結果から構成要素の決定木を用いて類似度行列を構築することによって，MDS が適用可能になる．各決定木が返す識別結果はデータの特徴量に依存しており，類似した特徴量を持つデータ間では各決定木の出力結果が類似する傾向にある．RF はモ

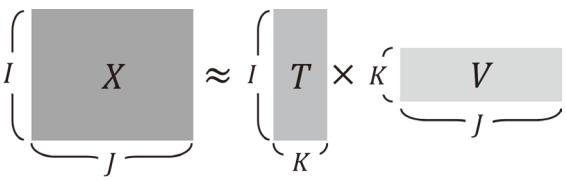

図 3　NMF の概略図

デル全体としての識別結果を出力する際に，各データに対して各決定木が識別結果を出力するため，ここでは，予測が一致した決定木の個数をデータ間の類似度として扱うことによって類似度行列が構築できる．一般に，このような処理は，RF の学習結果の内部モデルを視覚的に把握する際に利用される [17],[18]．この結果から，類似性を反映させたデータの可視化，すなわち，パテントマップの構築が可能である．一般的な MDS と異なる点は，類似度行列が，クラス分類の結果に基づいて構築されている点である．

ここでは，教師信号として扱われる特許の技術分野の予測マッピング対象であることから，一般的に用いられる文書間の類似性ではなく，クラスの情報がマッピングに反映されている．獲得されたマップ上で近傍に配置されるデータ同士の関係は，単に形態素の出現頻度の類似性に基づいたものではなく，構成要素である予測器によって同様な分類結果を返されたデータであること示している．周辺に異なる技術が多く配置している特許 (技術) は，(ノイズやモデルの不具合でない限り) たとえ分野が異なっていたとしても，他分野への適用可能性や，これまで知られていなかった新たな利用方法を構築できる可能性を秘めている．以上より，我々は，RF による MDS プロットを文書ベースのパテントマップへの利用は，技術の適用領域において新たな知見や発見となりうる情報を提示してくれると考えた．

3.2.2 カーネル密度推定

RF における MDS による 2 次元プロットでは，識別精度が高いほど，同じクラスを持つデータは近くに配置される．単に，ノイズ的なデータや適切に学習されなかったデータは，同クラスのデータの近くに配置されない可能性もあり，注目に値しないものであるにも関わらず，異なるクラスのデータの近くに配置されてしまう．本研究では，本質的に注目すべきデータを明確にするために，データが密集している領域に注目することにする．

このデータが密集している領域を特定するために，本研究では MDS のマッピングに対して，2 次元カーネル密度推定を施す．ここでは，マップ上の点 x に対する密度 $\hat{f}(x)$ は次式によって定義できる．

$$\hat{f}(x) = \frac{1}{nb}\sum_{j=1}^{n} K\left(\frac{x - x_j}{b}\right) \tag{5}$$

ここでは，$K()$ はカーネル関数を表しており，本研究ではガウシアン関数を採用している．また，b はバンド幅を表しており，一般的な b の推定値 $\hat{b}$ の計算方法は次式によるものであり，経験的に良好な結果が得られることが知られている [15].

$$\hat{b} = 1.06 \min(\hat{\sigma}, R/1.34) n^{-1/5} \tag{6}$$

ここでは，$\hat{\sigma}$ は標準偏差を，R は四分位範囲 IQR を，n はデータ数を表している．

この処理によって，データが集中している領域は高密度となりマップ上では（等高線として）高く表示される．これより，識別の傾向が視覚的に把握可能となり，知識抽出が可能になる．本研究では MDS において利用する類似度行列は RF による予測結果に基づいていることから，マップ上で近隣に配置されているデータは類似した予測結果が得られていることを表しており，類似した予測結果を持つデータほど密集する．このため，特に高密度の領域で，周辺のデータとクラスが異なるデータは高い確信度で識別を行ったのにも関わらず，同一のクラスに誤分類している可能性が高い．このようなデータは他分野において広く利用されているキーワードを多分に含んでいることから，他分野において適用の可能性が高い技術であると期待でき，獲得されたパテントマップ上では注目に値するといえる．

3.3 技術領域の分類に影響を及ぼす変数の特定

RF では，構成要素である決定木の平均精度や Gini 係数に基づいた変数重要度を算出し，この大小を比較することで識別に対して寄与した変数を特定することができる [17],[18]．すなわち，ここでいう分野の特定に影響を及ぼしている変数の特定が可能になる．また，本研究では大規模疎なデータである頻度行列を低次元密なデータに変換する際に NMF を利用している．先述のとおり，NMF はトピックとして次元縮約を行う方法であるため，軸の解釈が比較的容易であり，データの内容に関する解釈や理解支援が可能になる．さらに，このトピックの中で識別に高い影響を及ぼしているものを特定することで，獲得されたマップの内容やそこに影響を及ぼす要因の把握ができる．

4 実験（特許データへの適用）

ここでは，特許に関する実データを用いた分析を通じて提案法の挙動を確認する．

4.1 実 験 設 定

特許情報プラットフォーム J-PlatPat に公開されている特許情報において,「要約」および「特許請求の範囲」に含まれる文書データを対象として分析を行った（**注 3**）．技術内容が比較的把握しやすく挙動の確認を実施しやすいと思われる「ロボット」,「テキストマイニング」を分析における対象分野の一例とし，各

分野に対する 50 件を収集し，これに対して先述の分析を適用する．提案法における解析結果はデータへの依存が強く，対象の技術間に全く関係の無い領域であればデータ内に技術に関する共通項が全く含まれず，異分野への適用可能な特許が抽出できないことも起こりうる．

なお，**2** および **3** で示した変数とは単語の出現頻度の行列を次元縮約した結果に対応しており，ここでいう次元数 K はその総数である．また，クラスは分類対象であり，ここでは上記の「ロボット」と「テキストマイニング」に対応している．なお，本研究にて用いる RF に基づいたデータ間の類似度は，RF の構成要素である決定木による予測結果（出力）が一致した個数を利用している．比較対象として用いる一般的な文書間の類似度は，各文書データ間における単語の出現頻度ベクトルのなす角を θ とした際の $\cos\theta$ の値である．

分析の実行に際して，文書データから，形態素解析を通じて単語の頻度行列を作成した．本論は技術に関連する単語を分析対象としているため，抽出する素性は全て名詞とした．また，出現数が 5 回に満たない単語は削除した．ここでは,「ロボット」と「テキストマイニング」は分類対象の技術領域 (クラス) であるため,「ロボット」「テキスト」「マイニング」を素性から削除し，形態素から直接クラスを特定できない状況において提案する分析方法を適用した．なお，分析に用いた学習器の設定は表 1 に示すとおりである．

実験では，上記のデータに対して，まず提案法を適用し，分析結果から所望の知識抽出が実現できるかを確認する．また，本研究は RF の学習結果から知識を抽出することであるが，知識抽出の対象である，獲得されたモデルの妥当性を確認するために，識別精度の評価を行った．クローズドテスト・オープンテスト両方の評価を実施した．オープンテストとしては，1 個抜きの交差検証 (Leave One Out Cross Varidation, LOOCV) を行った．

4.2 実験結果

まず，図 4 および図 5 は変数の重要度を表している．前者は決定木の Gini 係数に基づくもので，後者は決定木の精度に基づくものである．これらは，横軸が NMF によって縮約された変数のインデクスに対応しており，縦軸が各分野を識別する際の変数の重要度を表している．

合成変数の内容を把握するために，図 4, 5 の両方の結果において，重要度が特に高かった（両図において上位 5 位以内に入っている）合成変数の単語の重み係数の一部を表 2 に示している．ここでは，それぞれの合成変数中において重み行列の対応する係数の大きさが上位 40 位以内に入っており，技術に関連すると考えられる単語を表示した．各単語の下の数値は重み行列の係数である．これらの単語の意味を考慮しつつ共通性を見出すことで，これらの合成変数の意味を解釈できる．

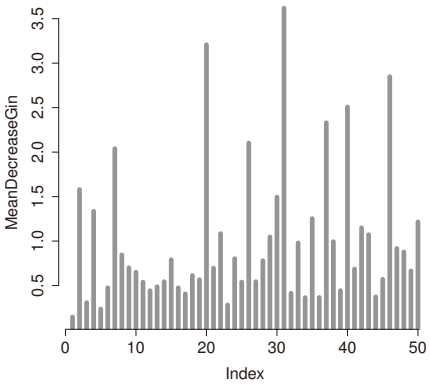

図 4　各変数の重要度 (Gini 係数)

表 1　分析におけるパラメータ設定

パラメータ	値
森のサイズ:B	50
各木で選択されるの変数の数	7
NMF の基底数:K	50
NMF の更新回数	100

実験の結果として獲得されたパテントマップは図 6 に示す通りである．比較対象として，広く利用されているコサイン類似度に基づいて作成した類似度行列に対して MDS を適用することによって作成したパテントマップを図 7 に示している．これらの図中の “T” で始まる黒色文字列は「テキストマイニング」関連の，“R” で始まる明灰色斜体の文字列は「ロボット」関連の特許に対応した記号である．また，図中の黒色実線の曲線はカーネル密度推定によって獲得された密度に関する等高線である．

表 3 は識別精度の評価である．クローズドテスト・オープンテスト共に森のサイズを $B = 5, 10, 50, 100$ とし，それぞれの設定において実施した結果を表している．

4.3 まとめと考察

表 2 においてそれぞれの変数における単語の内容に注目すると，以下のような解釈を与えることができる．

表 2　NMF の学習結果において重み係数が大きい単語の一例

変数	対応する単語とその重み									
V31	制御	移動	脚	駆動	接地	障害	位置	動作	衝突	歩行
	49.28	49.04	35.96	26.74	22.15	14.30	12.87	11.70	11.44	10.89
	設定	指令	軌道	車輪	旋回	関節	追随	計測	スイング	車体
	9.98	8.88	8.69	8.30	8.30	7.39	6.96	6.94	6.72	6.23
V20	光	クリーニング	集	位置	ミラー	装置	EUV	搬送	発生	回転
	86.79	40.27	70.96	69.19	52.38	43.16	31.96	26.01	18.49	13.89
	反応	クリーニングチャンバー	ガス	寿命	物質	汚染	交換	付着	対応	反射
	8.06	10.43	6.92	4.43	4.22	4.11	3.61	3.44	3.25	2.80
V46	情報	コミュニケーション	タイム	ライン	抽出	ログ	ライフ	商品	提供	関連
	142.97	36.73	28.29	27.80	23.93	22.97	22.97	21.35	19.89	19.19
	収集	データベース	検出	ユーザ	個人	端末	要素	蓄積	取得	DB
	17.91	11.61	16.65	16.19	13.29	12.53	10.69	7.53	6.34	5.41
V40	請求	コンテンツ	結果	情報	要因	クエリ	デバイス	ユーザ	記載	フォーマット
	58.47	28.72	48.53	45.99	45.84	43.30	28.93	28.23	27.97	21.08
	統計	レイティング	選択	アイテム	広告	スコア	販売	システム	ランキング	モジュール
	27.85	9.85	17.60	15.32	14.56	12.48	9.85	9.55	8.65	8.39
V37	情報	グルーピング	文書	データ	ユーザ	検索	オブジェクト	端末	登録	サーバ
	96.86	9.99	41.03	28.97	20.27	11.76	9.96	8.96	8.37	6.87
	管理	プレゼンテーション	ファイル	プログラム	送信	保存	グラフ	サマリー	企画	取得
	6.84	4.41	4.40	3.47	3.31	3.09	2.94	2.92	2.66	2.55

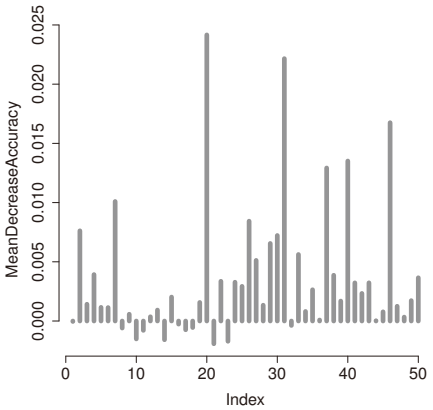

図 5　各変数の重要度 (Accurasy)

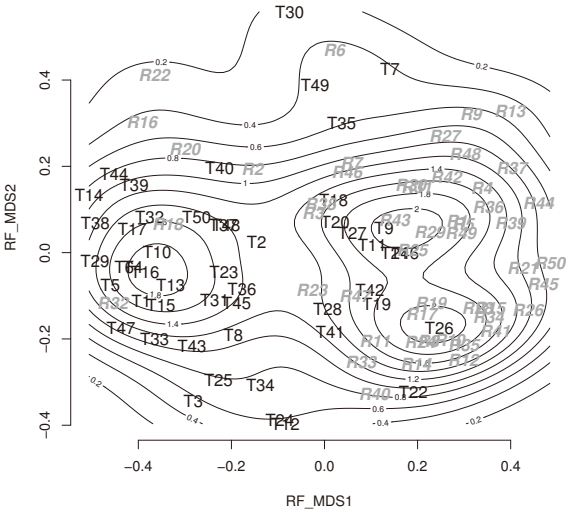

図 6　提案法によって獲得したパテントマップ

V31: ハードウェアに関するキーワードが多く，特に動作や制御に関連する単語に対する重みが強いため「駆動機器」に関する因子と解釈できる．

V20: ハードウェアに関するキーワードが多く，特に光を対象としたセンサ技術に関連する単語に対する重みが大きいため「光学機器」に関する因子と解釈できる．

V46: 情報処理技術に関するキーワードが多く，データベースを対象とした単語の重みが強いことから,「データ」に関する因子と解釈できる．

V40: 情報処理技術に関するキーワードが多く，評価や顧客との接点に関連する単語の重みが強いことから,「ビジネス」に関する因子と解釈できる．

V37: 情報処理技術に関するキーワードが多く，情報処理における基礎技術や通信技術に関連する用語の重みが強いことから,「情報・通信」に関する因子と解釈できる．

この結果から，共通の要因に合致した単語が多数出現しており技術情報に関する因子 (トピック) を適切に抽出できていることが確認できる．これは，一般に軸の解釈が難しい LSA や PCA に対して，このようなタスクにおいては優位な点であるといえる．

図 6 のマップの出力結果に目を移すと，マップの右側にロボット関連技術が，左側にテキストマイニング関連技術が集中していることが確認できる．カーネル密度の等高線に着目すると,「ロボット」と「テキスト

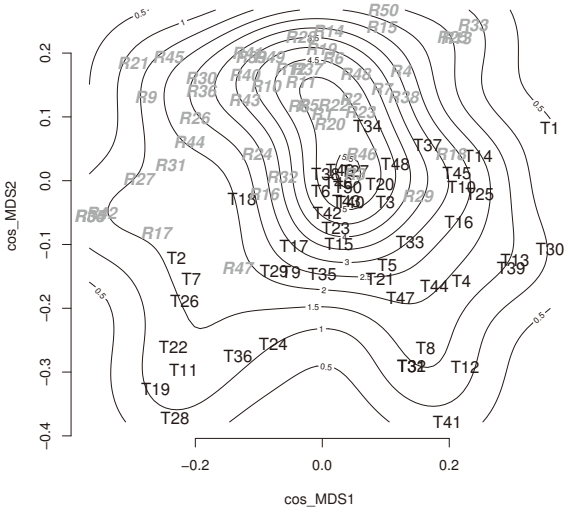

図 7　コサイン類似度に基づいた MDS によって獲得したパテントマップ

マイニング」の領域それぞれに対して高密度な領域が割り当てられていることがわかる．また，ロボットの領域はさらに上側と下側で二つの高密度領域に分かれている．これは，ロボット関連技術は機械や制御に関連する技術のみならず，情報技術や人間科学といった多様な要因から成る総合的な技術領域であることを示唆している．

上下二つに分かれているロボットの高密度領域の下側に注目すると，ロボット技術の中心部にテキストマイニングの技術 (T26) が配置されている．これは，テキストマイニングの技術であるにもかかわらず，ロボットへの応用が期待できる技術であると考えられる．特許内容の本文を確認したところ，音声認識に関連するキーワードを多く含むテキストマイニング技術であった．近年では，人間とコミュニケーションを取るロボットが社会的に注目されており，これはロボット技術として識別できるものであったと考えられる．先進企業の一部には特許戦略として，あえて適用が想定される領域の他分野に対して特許を出願するケースがあるという．この特許がそれにあたるかは定かではないが，このような結果を抽出できるアプローチは，自社の立ち位置の把握や他社との差別化，情報収集といった，企業における特許や技術に関する戦略策定において有益なツールとなりうると考えている．

次に，獲得された学習結果が妥当であるかを評価するために，提案法 (図 6) と一般的に用いられているコサイン類似度に基づいた多次元尺度構成法 (図 7) それぞれの結果として得られたパテントマップを視認によって比較する．文書間の類似度をコサイン類似度に基づいて算出した多次元尺度構成法では，技術領域がきれいに分かれて表示されているが，データの高密度な領域が一か所に集中しており，適用可能な領域の把握には適さないといえる．これは，RF の構成要素の分類結果を類似度として構築した本提案が目的に合致していることを支持する結果となった．最後に，データの当てはまり具合を評価すると，表 3 の結果より，高い識別精度が確認できることから，獲得されたマップには一定の信頼を置けるといえる．

表 3　予測精度の評価

森のサイズ:B	$B=5$	$B=10$	$B=50$	$B=100$
オープンテスト	75%	84%	85%	87%
クローズドテスト	98%	100%	100%	100%

5 おわりに

本研究では，特許の文書データ分析において，他分野との関連性や分野間の差異を示す要因を把握することを目指して，NMF により縮約したデータに対して RF 適用した．RF の学習結果に基づいた MDS によるマッピングを通じ，その内部構造を理解することでパテントマップを構築した．これに対してカーネル密度推定を適用することで，技術領域 (クラス) 間の関係における特許に関する適用領域の明確な可視化を可能とし，特許に関するデータの視覚的な理解支援を実現できたと考えている．

将来的には，ICT のさらなる伸展に伴い，技術情報は様々な領域に跨って複雑化・高度化すると容易に予想できる．特に近年では，社会的インパクトが大きい技術の開発やビジネスにおける成功を目指す上では，技術的に高度な発明のみでは不十分で，事業化への視点を伴わせた技術開発が重要であることが指摘されている．このことから，技術と経営を連動させることを目的とした，技術に関するデータ分析に対するニーズはさらに高まると考えられる．以上より，本論のアプローチのように機械学習や計算機統計学をベースにした特許や技術情報に関する分析技術およびマイニング技術の重要性は，今後さらに高まるのではないかと考えている．

今後の課題として，さらなる精度向上や精緻な知識抽出を目指す上で，単語間の関連性を考慮した分析（ネットワーク分析，係り受け解析）や文書以外の情報とのデータ統合を進めていくことが必要であると考えている．

注

(1) 選択する変数の数は $k = \lfloor\sqrt{d}\rfloor$ が推奨されている．本研究においてもこれに基づいて変数選択を実施した．

(2) NMF の更新式は近似された行列とデータ間の距離に基づいて様々なものが提案されているが，本研究では，広く用いられているユークリッド距離に基づいた更新式を利用している．

(3) 特許情報プラットフォーム J-PlatPat
https://www.j-platpat.inpit.go.jp/

謝辞

本研究の実施にあたり，特許情報プラットフォーム J-PlatPat におけるサービスを利用させて頂きました．ここに記して感謝いたします．また，有益なコメントを頂きました査読者・編集委員の皆様に御礼申し上げます．

参 考 文 献

[1] 野崎篤志：「特許情報分析とパテントマップ作成入門」，発明推進協会 (2011)

[2] 奥村学 (監修), 藤井敦, 谷川英和, 岩山真, 難波英嗣, 山本幹雄, 内山将夫：「特許情報処理：言語処理的アプローチ」，コロナ社 (2012)

[3] 酒井浩之, 野中尋史, 増山繁：“特許明細書からの技術課題情報の抽出”, 人工知能学会論文誌，Vol. 24, No. 6, pp. 531-540 (2009)

[4] 坂地泰紀，野中尋史，酒井浩之，増山繁：“Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法”, 電子情報通信学会論文誌 D，Vol. J93-D, No. 6, pp. 472-755 (2010)

[5] 西山莉紗, 竹内広宣, 渡辺日出雄, 那須川哲哉：“新技術が持つ特徴に注目した技術調査支援ツール”, 人工知能学会論文誌，Vol. 24, No. 6, pp. 541-548 (2009)

[6] 菰田文男：“「単語セット」の作成と進化に基づくテキストマイニング手法-MOT（技術経営）のためのテキストデータ解析を事例として”, 情報管理, Vol. 54, No. 9, pp. 568-578 (2011)

[7] 渡部勇：“テキストマイニング技術による公知例調査の支援”, 日本ロボット学会誌，Vol. 22, No. 3, pp. 304-307 (2004)

[8] Segev, A. and Kantola, J.: “Identification of Trends from Patents Using Self-organizing Maps,” *Expert Syst. Appl.*, Vol. 39, No. 18, pp. 13235-13242 (2012)

[9] Trappey, A.J.C., Trappey, C.V., Wu, C. and Lin, C.：“A Patent Quality Analysis for Innovative Technology and Product Development,” *Adv. Eng. Inform.*, Vol. 26, No. 1, pp. 26-34 (2012)

[10] Wu, J., Chang, P., Tsao, C. and Fan, C.：“A Patent Quality Analysis and Classification System Using Self-organizing Maps with Support Vector Machine,” *Appl. Soft Comput.*, Vol. 41, pp. 305-316 (2016)

[11] Breiman, L.: “Random Forests,” *Mach. Learn.*, Vol. 45, pp. 5-32 (2001)

[12] Efron, B. and Tibshirani, R.: An Introduction to the Bootstrap, Chapman and Hall, New York (1993)

[13] Lee, D.D. and Seung, H.S.: “Learning the Parts of Objects by Non-negative Matrix Factorization,” *Nature*, Vol. 401, pp. 788-791 (1999)

[14] 豊田裕貴, 菰田文男：「特許情報のテキストマイニング-技術経営のパラダイム転換」，ミネルヴァ書房 (2011)

[15] Venables, W. V. and Ripley, B. D.: Modern Applied Statistics with S (Fourth edition), Springer, (2002)

[16] 丹羽清：「技術経営論」，東京大学出版会 (2006)

[17] 下川敏雄, 杉元知之, 後藤昌司, 金明哲 (編)：「樹木構造接近法」，共立出版 (2013)

[18] Berk, R. A.: Statistical Learning from a Regression Perspective (2nd ed.), Springer (2016)