機械学習を用いた効率的な特許調査方法

Effective patent search methods using Machine Learning

花王株式会社　知的財産部／アジア特許情報研究会　**安藤　俊幸**

1985 年現花王株式会社入社、研究開発に従事
1999 年研究所の特許調査担当（新規プロジェクト）
2009 年より現職
2011 年よりアジア特許情報研究会所属

1 はじめに

近年、グーグルのネコを認識する人工知能（2012年）、自動車の自動運転、囲碁における人工知能の世界トップ棋士に勝利等人工知能に関する話題に事欠かない状況であり、ディープラーニング（深層学習）をベースにした第３次人工知能ブームになっている[1)]。また内閣府の知的財産戦略推進事務局による「知的財産推進計画 2016」[2)] においては IoT（Internet of Things：モノのインターネット）、ビッグデータ、人工知能は第４次産業革命という文脈で注目されている。日本特許庁の「平成 28 年度 人工知能技術を活用した特許行政事務の高度化・効率化実証的研究事業」[3)] の公募が行われ外部有識者による選定委員会での審査を経て株式会社エヌ・ティ・ティ・データ経営研究所が委託事業者に選定されている。民間の特許調査の分野においても 2015 年の特許情報フェアで UBIC 社（現株式会社 FRONTEO）の講演「人工知能による特許調査・分析業務の効率化と最新事例」があり会場を変更するほどの大盛況であった。ただ一般論として「人工知能」と聞いてアニメや SF に登場する例えば「ドラえもん」のような何でも希望を叶えてくれるロボットのような存在（人工知能分野で言うところの強い AI）をイメージして期待値が上がっている一面もあるようにも思われる。この期待は逆に言うと現状の特許検索データベースや特許調査用の種々のツールに潜在的な情報要求、いわゆる直感的要求（現状に満足していないことは認識しているが、それを具体的に言語化してうまく説明できない状態）があるように思われる。

現状の「人工知能」は人工知能学会のページ「What's AI　人工知能研究」[4)] を参考にすると機械学習、自然言語処理、情報検索、データマイニング、画像認識、音声認識等の様々な情報処理技術から成る複合技術であることが分かる。なお人工知能の特許調査・分析・解析への応用動向に関しては本寄稿集の桐山氏の論文で詳しく解説されている。

本稿では特許調査・解析の実務に自分の手を動かして実際に試せる特許調査の効率化手法を主に機械学習の観点から検討した。

2 特許調査への機械学習の応用の目的

機械学習の特許調査への応用の目的として下記２種類の特許調査をベースに目的を設定する。

①技術動向調査

データから規則性を学習する機械学習の技術を用いて膨大な特許情報から技術動向を効率的に把握する。全体像が直感的に把握できて関心がある特許公報にインタラクティブ（対話的）にアクセスできるような俯瞰・可視化とインタラクティブ操作機能を備えたツールが理想的である。英語、日本語、中国語で解析可能で日本以外の特許も F タームを使用するような感覚で処理できると便利である。

②先行技術調査

機械学習の観点では教師データが少なくても効率的に学習して再現率と精度を両立可能な調査手法。特許検索

の観点では検索漏れを少なくするように網羅性を重視した検索母集団を作成し精度を重視したスクリーニングを行い調査目的に適合したスコア付けを行う調査手法を目的とする。更に適合した部分を例えば段落単位で提示する。

3 機械学習とは

機械学習とは、データから学習した結果をもとに、新たなデータに対して判定や予測を行うことである。

機械学習は検索エンジン、医療診断、スパムメールの検出、金融市場の予測、DNA 配列の分類、音声認識や文字認識などのパターン認識、ゲーム戦略、ロボット、など幅広い分野で用いられている。応用分野の特性に応じて学習手法も適切に選択する必要があり、様々な手法が提案されている。

機械学習とデータマイニングは類似部分が大きく、技法も同じなので混同されることが多いが、次のように定義できる。

機械学習の目的は、訓練データから学んだ「既知」の特徴に基づく予測である。

データマイニングの目的は、それまで「未知」だったデータの特徴を発見することである。[5)]

本稿ではデータマイニングツールを使用して機械学習を用いて特許調査の効率化検討を行う。

4 機械学習ツール

機械学習ツールの紹介や選択方法等機械学習の入門向け参考文献としては「データサイエンティスト養成読本機械学習入門編」[6)] がある。

プログラミング言語系ツール

・R 言語（統計解析向けのプログラミング言語）

・Python（汎用スクリプト言語）

プログラミングを厭わない人には機械学習に強いプログラミング言語として R 言語、Python[7), 8)] が有名である。

「R によるデータサイエンス - データ解析の基礎から最新手法まで」[9)] は機械学習に関する記述もあるがデータサイエンスとして幅広い手法を扱っている。テキストマイニングを学ぶ上で基礎的な「テキストデータの統計科学入門」[10)] の一部内容を加えて Web[11)] 上で閲覧できる。

Python には機械学習のためのライブラリが各種そろっておりプログラミング言語として生産性が高く習得しやすいと言われている。

プログラミング言語は敷居が高いと言う人には GUI ベースで操作できる Weka[12)] がある。Weka（Waikato Environment for Knowledge Analysis）はニュージーランドのワイカト大学で開発された機械学習ソフトウェアで、Java で書かれている。GNU General Public License でライセンスされているフリーソフトウェアである [13), 14)]。Weka はデータマイニングツールとしても使用できる。

KH Coder[15)] はテキストマイニング用のフリーツールであり文書を分類したり、R を用いた多変量解析と可視化機能を備える。「社会調査のための計量テキスト分析」[16)] は KH Coder の作者が書いた著作で KH Coder のチュートリアル・マニュアルに加えて、分析の考え方、KH Coder の主な機能、有効性の検証、応用事例、社会調査での活用の展望などがまとめられている。KH Coder は表現 A が文書中にあれば、事柄 A が出現していたと見なして文書に A と分類を付与するコーディングルールによる分類に加えて人間がいくつかの文書を分類して「見本」を示せば、そこから分類の基準を学習して他の文書を自動的に分類する「ベイズ学習による分類」機能も有している。

「人工知能による文書分類」[17)] では文書分類で SVM（Support Vector Machine）、k 近傍法、ロジスティック回帰等が良く使われる手法であり、特に k 近傍法、ナイーブベイズ分類について分かりやすく解説されている。

以上、フリーのツールと自分で使うことに主眼を置いた文献を挙げたがもっと簡単に特許に対して適用してみたいという人には有償であるが NTT データ数理システムの下記３種類のツールがある（１カ月の試用も可能）。

①特許情報分析ツール：Patent Mining eXpress（PMX）[18)]

Python 等で作成された Web ブラウザから操作する特許情報分析の専用ツールである。書誌、抽出キーワードの統計機能等良く使う解析項目はメニュー化されている。

②テキストマイニングツール：Text Mining Studio（TMS）[19]

汎用のテキストマイニングツールである。「特許情報のテキストマイニング」[20] にツールと特許情報の解析事例の紹介がある。

③汎用データマイニングシステム：Visual Mining Studio（VMS）[21]

Visual Mining Studio は簡単な GUI 操作で本格的なデータマイニングが行えるツールである。データの前処理から、マイニング処理、他アプリケーションとの連携機能を備え、さらにその結果をグラフィカルな表示で表現することができる。TMS とシームレスに連携して TMS によるテキストマイニングの出力を更に高度にマイニングしたり、機械学習により教師データありの分類や教師データなしのクラスタリング処理等が行える。

5 特許調査への機械学習適応時の留意点

機械学習はその本質上データが重要である。ロジックが正しくてもデータが間違っていると正しい答えは期待できない。機械学習に限らず特許調査において特許データベースのデータ収録状況は特許調査の品質に直接的に影響するので極めて重要である。

・収録国

・レコード（公報）収録、収録率、タイムラグ

・フィールド（項目）の収録

新興国ではこれら各種収録状況を調べること自体が非常に困難で忍耐と「気合い」を要する。これら収録状況を把握しておかないと調査の質を担保出来ない。アジア特許情報研究会のホームページにこれまでの研究成果が一覧にまとめて公開されている [22]。

後述する Questel 社の特許データベース Orbit.com[23] のコンセプト（テキストマイニング的に抽出した英語のテクニカルワード）情報が中国特許には機械翻訳の英文より抽出されて付与されているが日本特許には英語公報のファミリーがあるものを除いてコンセプトが付与されていない（今年中には付与される見込みと聞いている）。Orbit のデータを解析する Analysis module のランドスケープマップ等コンセプト情報を用いる解析には要注意である。

サイテーション（引用 / 被引用）データ未収録

海外データベースベンダーの特許データベースに収録されている日本特許のサイテーション情報は数 10 パーセントのオーダーで収録されていないデータベースが大多数である。きちんと収録されているデータベースはレアケースである。これは欧州特許庁の Espacenet 等も例外ではない。日本のベンダーのデータベースと比較すると一目瞭然である。海外製データベースで日本特許のサイテーション情報を利用した解析を行う場合は要注意である。

国別 IPC 付与状況

各国の産業の発展度合いや文化的な背景等様々な要因で IPC の付与状況は異なる。

IPC の付与ロジックミスの事例

ファミリー型の商用特許データベースでバージョンアップ後 IPC 検索においてヒット件数が場合により 1 万件オーダーで少なくなる現象が発生したことがある。過去のダウンロードデータおよび他の商用データベースと IPC の付与状況を比較したところ IPC の版により IPC 付与分布が異なっていた。特に IPC7 版が圧倒的に少なくなっていた。2006 年に IPC7 版から IPC8 版で大幅に改正された分野、例えば化粧品（7 版：A61K7、8 版：A61K8）で解析すると良く分かる。データベースのバージョンアップ時に IPC7 版の取扱いロジックを誤ったと推定し網羅性（再現率）と精度（適合率）の両方で問題である旨ベンダーに指摘した。後日談で対象レコード数千万件を修正したとのことであった。

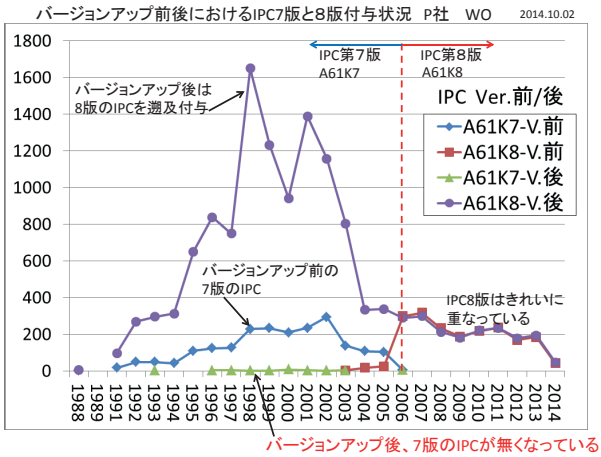

図 1　データベースの IPC 付与エラー解析例

デジタル化以前の日本の公報や海外のイメージ公報は OCR でテキスト化されてデータベースに収録されてい

るがデータベースにより OCR の品質が異なっており書誌事項を始めテキストデータの品質も様々である。

データベースにより公報番号表記形式が異なっているのはごく普通であり同じデータベースでも公報データと経過情報では公報番号表記が異なっている場合もある。権利調査等で経過情報をマージして使用する場合には注意を要する。

出願人の記述する言葉が異なるキーワードの異表記や機械翻訳の誤訳、文字コードの違いによる文字化け等落とし穴になりそうなポイントは多数存在する。事前に分かっているものは対処できるが結果を検証(答え合わせ)して想定外の誤りが無いか確認することは重要である。

昔からコンピュータの世界では garbage in, garbage out (GIGO)[24]:直訳すると「ゴミを入れると、ゴミが出てくる」と言われていたが解析の分野で「収集・蓄積したデータが不正だと、分析結果も不正になる」という意味—すなわちデータクレンジングの重要性を示す言葉として使われることが増えてきている。

6 技術動向調査での機械学習の利用事例

技術動向調査事例として特許庁の「人工知能」の動向調査[25]を参考に Orbit.com の下記検索式 の 22457 ファミリーを Analysis module を使用して解析した（図2～図8）。

（G06N）/IPC/CPC AND PD=2006-01-01:2016-06-30

G06N は人工知能分野の IPC である。PD は公開日である。この集合の優先国:33、発行国:46、出願数ベースで 57778 件であった（共に EP、WO を含む）。

表1　優先国と発行国（上位 10）

発行国＼最初の優先国	CN	DE	EP	GB	JP	KR	RU	TW	US	WO
AU	1	14	40	48	40	6	1	1	881	50
CA	4	18	40	37	39	5	0	0	1075	37
CN	4227	74	97	52	435	63	3	23	1355	139
DE	4	329	59	24	91	10	0	0	311	29
EP	63	161	401	151	366	65	4	6	2168	185
IN	7	22	42	28	30	8	1	1	540	34
JP	47	52	120	76	2718	60	0	12	1315	110
KR	14	17	42	20	169	472	1	6	619	43
US	210	197	354	233	1321	322	11	118	11219	353
WO	127	196	199	179	518	99	16	0	3578	469

表1に最初の優先国と発行国の上位 10 位のマトリックスを示す。列が最初の優先国で行が発行国である。最初の優先国で発明がなされ、発行国でどこの国に出願されているか把握できる。WO を除くと US が圧倒しており次に CN、JP と続く。公報の使用言語の順番は英語、中国語、日本語になる。この表から CN（中国語）、JP（日本語）しかない公報が相当数あることが推定される。

図2は予め定められた IPC に基づいて公報をクラス分類している。技術領域として Computer technology に集中している。また応用特許が非常に幅広い分野に出願されている。各 Technology domain（ヘキサゴン：六角形）の位置は予め決まっており変わることはない。各ヘキサゴンに振り分けられた特許

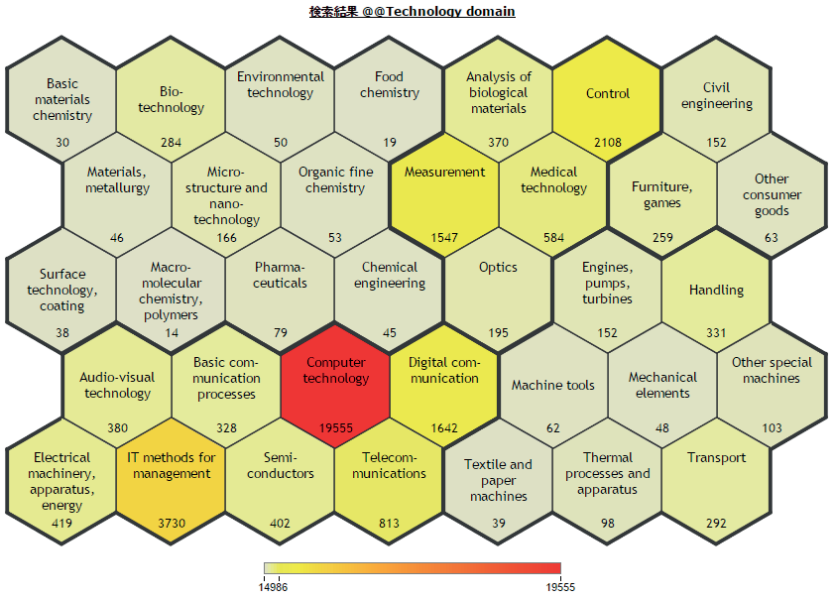

図2　Technology domain のヘキサゴンチャート

（ファミリー）数に比例して色が変わる。Technology domain のヘキサゴンチャートは一層構造である。特定のヘキサゴンをクリックしてフィルタリングすることは可能である。これに対して後述する FoamTree は二層構造でクラスタリングアルゴリズムが用いられており表示するたびにポリゴン（多角形）の位置、形状、配色が変化する。

商用データベースの解析ツールや汎用のテキストマイニング、データマイニングツールでは「機械学習」はあまり意識しないがマイニングの技法として機械学習の教師データ無しのクラスタリングはよく使われている。

図３　コンセプトのタグクラウド（上位 100）

コンセプトのタグクラウドは対象母集団（公報１件でも表示可能）中のコンセプトの頻度情報を使用して高頻度のコンセプトを大きなフォントで表示している。表示位置には格別な意味はない。

表２にコンセプト上位 30、表３に出願人上位 30 を示す。出願人は統制（名寄せ）していない。

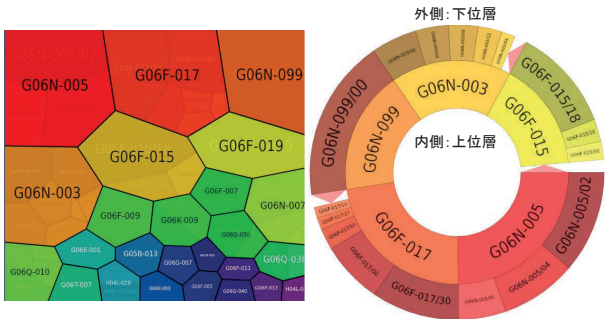

図４　IPC の FoamTree とドーナツチャート

図４の IPC の FoamTree とドーナツチャートは共に内部的には２層になっており上位層（概念）の下に下位層（概念）がある。IPC 自体にも階層関係がありドーナツチャートでは階層関係が明らかである。FoamTree とドーナツチャートは同じ色でカラーマッピングされて

表２　コンセプト上位 30

No.	コンセプト	@@Value
1	ALGORITHM	6319
2	DATABASE	5428
3	PROBABILITY	4769
4	RULE	4380
5	COMPUTER	3704
6	NEURAL NETWORK	3645
7	DECISION	3454
8	SUBSET	3453
9	VECTOR	3453
10	CLASSIFICATION	3260
11	SEARCH	3239
12	TRAINING	3033
13	BEHAVIOR	3013
14	CATEGORY	2948
15	KNOWLEDGE	2847
16	ATTRIBUTE	2825
17	LEARNING	2795
18	ITERATION	2727
19	OPTIMIZATION	2693
20	TRAINING DATA	2549
21	COMPUTER READABLE MEDIUM	2297
22	COMPUTER READABLE STORAGE	2259
23	COMPUTER PROGRAM PRODUCT	2229
24	READABLE STORAGE	1977
25	COMPUTING DEVICE	1882
26	GENETIC ALGORITHM	1873
27	SIMILARITY	1824
28	LEARNING ALGORITHM	1817
29	NEURON	1720
30	COMPUTING ENVIRONMENT	1665

表３　出願人上位 30

No.	出願人	@@Value
1	IBM	1347
2	MICROSOFT TECHNOLOGY LICENSING	724
3	NEC	337
4	SONY	334
5	GOOGLE	328
6	FUJITSU	275
7	SIEMENS	254
8	HEWLETT PACKARD	249
9	NIPPON TELEGRAPH & TELEPHONE	226
10	TOSHIBA	194
11	QUALCOMM	186
12	STATE GRID CORPORATION OF CHINA	182
13	YAHOO	177
14	SAP	171
15	SAMSUNG ELECTRONICS	162
16	ORACLE	137
17	HITACHI	132
18	MICROSOFT	132
19	XEROX	131
20	INTEL	117
21	ZHEJIANG UNIVERSITY	111
22	GENERAL ELECTRIC	110
23	BEIHANG UNIVERSITY	109
24	BEIJING UNIVERSITY OF TECHNOLOGY	102
25	XIDIAN UNIVERSITY	101
26	CANON	100
27	MITSUBISHI ELECTRIC	97
28	TSINGHUA UNIVERSITY	90
29	DWAVE SYSTEM	81
30	KOREA ELECTRONICS TELECOMM	79

いるので両方の図を比べて見ることで階層関係を把握できる。

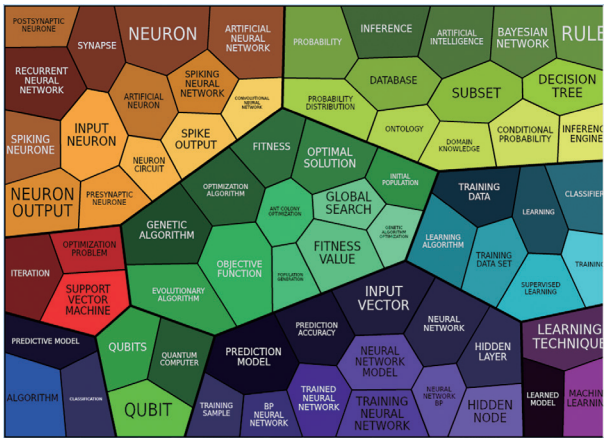

図５　FoamTree: interactive Voronoi treemap

図５の FoamTree は、JavaScript の革新的なレイアウトアルゴリズムとアニメーションでツリーマップとして可視化したものである。このような文書のグループ、ネットワークドメインやサイトマップなどの階層データの理解を助ける。内部アルゴリズム的にはコンセプトを使用して公報のクラスタリングを行っていると考えられる。２万件を超えるこの事例でも表示まで約 20 秒と非常に高速なアルゴリズムである。各クラスター（ポリゴン：多角形）をマウスでクリックするとそのクラスターに属する公報がフィルタリングされる。コンセプトのクラスタリングである図５、図６共に内部的には２層になっており上位層（概念）の下に下位層（概念）がある。IPC の FoamTree とドーナツチャートと同様である。図６は階層関係が明らかである。

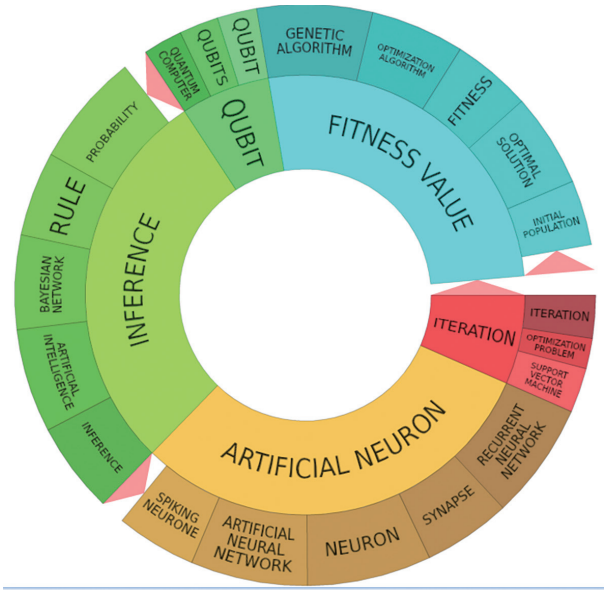

図６　コンセプトのドーナツチャート

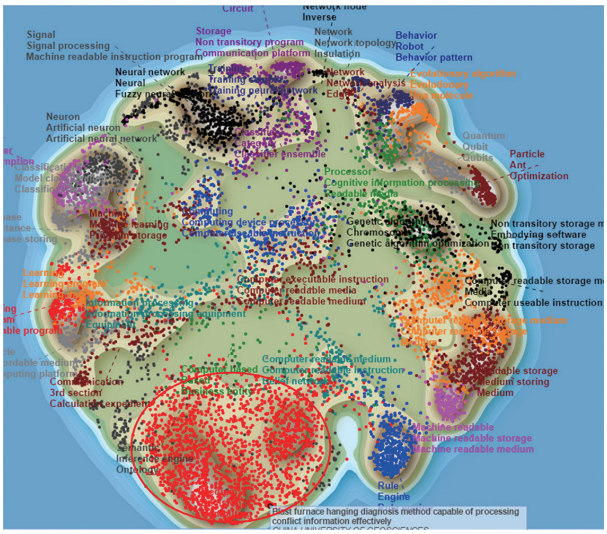

図７　Landscape map

Landscape map はコンセプト情報を用いて公報間の類似度（距離）を計算して俯瞰可視化している。色付きのドットは各公報である。同じクラスターには同じ色が割り当てられている。この Landscape map の処理時間は約 30 分であった。Landscape map からマウス操作でインタラクティブに任意の領域（例えば赤の楕円）あるいはクラスターを選択してサブ集合を解析したり、個々の公報にアクセスできる。個々の公報の出願人別にカラーマッピングを指定することもできる。

図８は出願人間の引用情報をネットワークグラフとして表示したものである。最少ノード数 20、最少リンク数 20 に限定して主要な出願人の引用関係を描かせている。

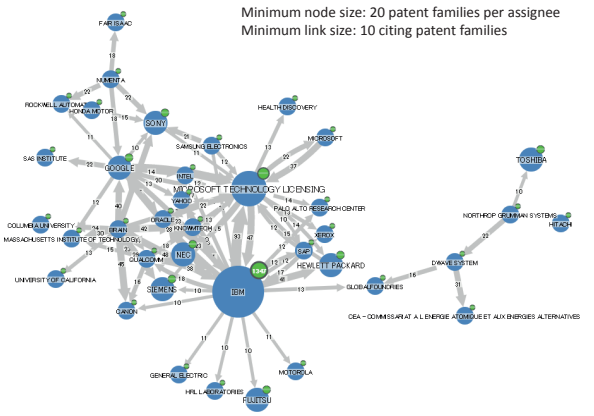

図８　出願人引用の Node chart

図９、図 10 は NRI サイバーパテントデスク２を使用して IPC:G06N、公開日：20060101：20160630、公報種別：A、T、S の 1612 件を母集団として NTT データ数理システムの特許情報分析ツー

ル：Patent Mining eXpress（PMX）を使用して解析したものである。

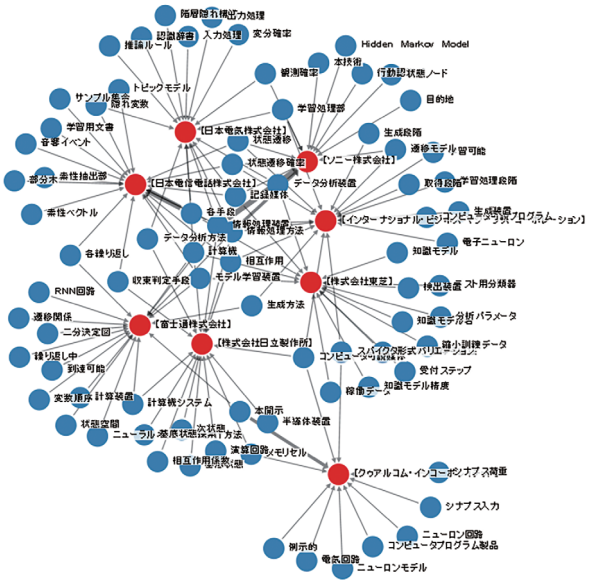

図９　技術特徴ネットワークグラフ

赤いノードが出願人で周りに抽出されたキーワードが表示されている。各出願人の特徴が読み取れる。

図 10 は PMX のデフォルト条件での課題と解決手段のキーワード抽出結果である。PMX には公報間位置関係を示すポジショニングマップ機能もある [26]。

市販のパテントマップソフト、例えばインパテック社のパテントマップ EXZ を使用すれば Orbit.com のデータや各種日本特許データベースのデータを取り込み書誌事項や抽出キーワードで各種統計処理マップを描かせることは極めて容易である。ただし課題と解決手段の関係をマップ化しようとすると自分でキーワードを選択する必要がある。キーワードの選択は必ずしも短所ではなくノイズを除去したり、課題と解決手段を技術内容に沿った形で並べたり、関心のある分野を抽出したりとメリットも多い。

クラスタリングやネットワーク表示による俯瞰可視化ツールを用いるとあまり（教師データなし）機械学習を意識しないでもマップ表示は容易である。ただし内部処理をある程度分かっていないと解析結果の解釈や説明が難しい。筆者はテキストマイニングによる特許調査の効率化を検討してきた [27)、28)]。機械学習の理解はテキストマイニングとも内部処理は共通な部分も多いのでお互いに有用である。

広義での分類に関する言葉としてグルーピング（grouping）、セパレーション（separetion）、セグメンテーション（segmentation）、カテゴリゼーション（categorization）、クラシフィケーション（classfication）等がある。特許庁の特許出願技術動向調査報告 [25)] で行われている人間が読み込み体系的にクラス分類するのはクラシフィケーションである。機械学習の観点からは下記のように分けられる。

クラスタリング：目的変数のない（教師なしの）場合

クラス分類：目的変数のある（教師ありの）場合

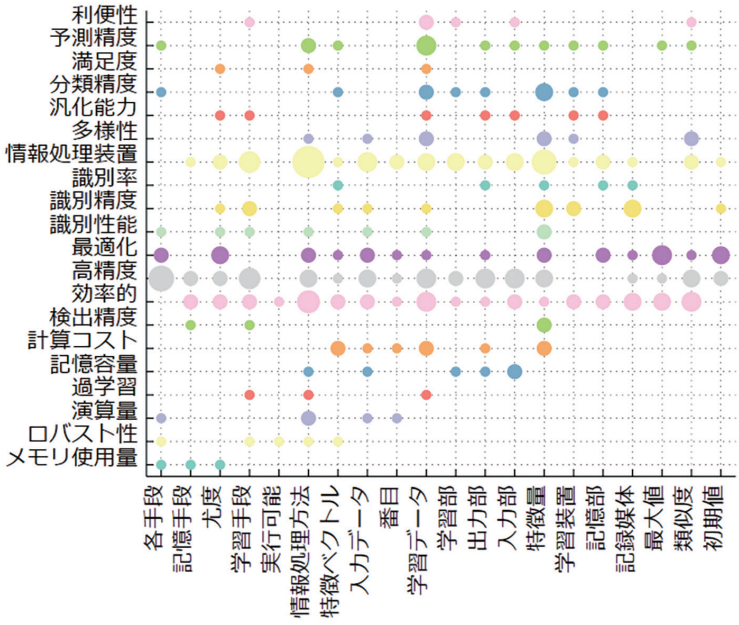

図 10　課題と解決手段の関係

次に教師データありの機械学習の先行技術調査への応用検討を述べる。

7 先行技術調査への機械学習の応用

（株）FRONTEO（旧 UBIC）の人工知能による特許調査は先行技術調査や無効資料調査の効率化で注目されていることと思われる。FRONTEO の人工知能を利用した PATENT EXPLORER[29] に関しては機関投資家向けセミナー資料[30]（2014年3月6日）が参考になる。セミナー資料によると人工知能のベースになっている技術分野は機械学習、行動情報科学、自然言語処理である。技術的な特徴として下記①～③が挙げられている。

①伝達情報量（transinformation）

②文書のスコアリング

③文書のコーディング＝自動仕分け

FRONTEO の独自技術として Landscaping とネーミングされた機械学習手法であり学習データを多面的に評価する学習機能を持ち、データが少数であっても効率的な学習が可能。計算コストが少ないとされる。

（従来からある Orbit.com の Landscape map とは別物なので混同しないよう注意が必要である。）

PATENT EXPLORER では上記の特徴に加え段落検索機能が加わっている。

伝達情報量は相互情報量（Mutual information）[31]とも呼ばれている。

PATENT EXPLORER 関連情報と「これからの特実検索システムの探求」[32] を参考にして先行技術調査の効率化検討を進めている。主な検討ツールとしてNTTデータ数理システムの②テキストマイニングツール：Text Mining Studio（TMS）と③汎用データマイニングシステム：Visual Mining Studio（VMS）を使用して教師あり学習を検討した。検討技術分野は即席麺の直近10年から母集団を作成しその中から教師データとなる公報を選択した。教師データとして審査官が拒絶理由に採用した文献の中から拒絶理由の条文コード[33] を利用して新規性（カテゴリーX文献）、進歩性（カテゴリーY文献）欠如に該当する文献を使用した。条文コードだけでは新規性と進歩性を完全に分離することはできないのでさらに拒絶理由通知書で適用条文を確認した。拒絶理由通知書を吟味して教師データとしてふさわしい公報を選択した。

表4にTMSの分かち書き出力例を示す。行IDが公報固有のユニークなキー（例えば出願番号）に相当する。文章IDは公報中の文に相当する。教師データとして与えた公報に近い公報を予測することを検討しているが今後、文単位あるいは段落単位で判別ことを検討する予定である。

表4　TMSのテキストマイニング分かち書き出力例

ファイルID	行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳細	係り先	述語属性	関係子
1	2	1	1	請求項	請求項	請求項	名詞	一般	2	なし	限定
1	2	1	2	1	1	1	名詞	数	3	なし	限定
1	2	1	3	炭酸カルシウム、	炭酸カルシウム	炭酸カルシウム	名詞	一般	10	なし	状況
1	2	1	4	燐酸カルシウム	燐酸カルシウム	燐酸カルシウム	名詞	一般	10	なし	状況
1	2	1	5	以下、	以下	以下	名詞	副詞可能	7	なし	状況
1	2	1	6	カルシウム剤と	カルシウム剤	カルシウム剤	名詞	一般	7	なし	現象
1	2	1	7	記す	記す	記す	動詞	自立	4	なし	注釈
1	2	1	8	及び	及び	及び	接続詞		9	なし	状況
1	2	1	9	ドロマイトから	ドロマイト	ドロマイト	名詞	一般	10	なし	状況
1	2	1	10	なる	なる	なる	動詞	自立	11	なし	限定
1	2	1	11	群から	群	群	名詞	一般	12	なし	状況
1	2	1	12	選ばれた	選ぶ	選ぶ	動詞	自立	20	なし	限定
1	2	1	13	少なくとも	少なくとも	少なくとも	副詞	一般	20	なし	状況
1	2	1	14	1種100重量	1種100重量	1種100重量	名詞	数	16	なし	限定
1	2	1	15	A	A	A	名詞	一般	14	なし	注釈
1	2	1	16	部に対し、	部	部	名詞	一般	20	なし	限定
1	2	1	17	加工デンプンを	加工デンプン	加工デンプン	名詞	一般	20	なし	現象
1	2	1	18	B	B	B	名詞	一般	17	なし	注釈
1	2	1	19	0.1～80重量	0.1～80 0重量	0.1～80重量	名詞	数	20	なし	限定
1	2	1	20	部含有させて	部含有	部含有	名詞	サ変接続	21	なし	状況
1	2	1	21	なることを	なる	なる	動詞	自立	22	なし	現象
1	2	1	22	特徴とする	特徴	特徴	名詞	一般	23	なし	限定
1	2	1	23	食品添加剤スラリー組成物	食品添加剤スラリー組成物	食品添加剤スラリー組成物	名詞	サ変接続	-1	なし	なし

【請求項1】炭酸カルシウム、燐酸カルシウム（以下、カルシウム剤と記す）及びドロマイトからなる群から選ばれた少なくとも1種（A）100重量部に対し、加工デンプン（B）を0.1～80重量部含有させてなることを特徴とする食品添加剤スラリー組成物。

TMS は９月１日にバージョンアップして Text Mining Studio 6.0[34] になった。表４の TMS の分かち書き出力例はバージョンアップ前のものである。バージョンアップではより高精度・高機能な日本語解析処理の実現を目指し、日本語解析エンジンの全面刷新が行われている。それに伴い、分かち書き結果における以下の点が大きく変更されている。

・品詞体系の刷新
・「述語属性」を「態度表現」に名称変更し、内容の刷新と再編
・自動連結処理の見直し
・見出し語から括弧が除外されて現れなくなる問題を解消

図 11 は Visual Mining Studio（VMS）と Visual R Platform の両方を担う Visual Analytics Platform である。VMS は各種の機械学習の手法を使うことができる。左側の Object Browser のウィンドウ部に各種処理や機能をオブジェクトとして系統的にまとめてアイコン化されている。必要な処理や機能のアイコンをマウスで選択してプロジェクトウィンドウ部において各アイコン間を線で繋いで必要な処理を選択することでプログラミング不要の GUI 操作で各種機械学習（モデリング）やクラスタ分析、アソシエーション分析、多変量解析等を非常に簡単に行うことが可能である。Visual R Platform で同様に GUI 操作で R 言語の処理を実行できる。

教師データありの機械学習の種類はまず対話型モデル学習から検討を始めた。対話型モデル学習は下記の特徴を有している。対話型モデル分析機能により、一部のデータにしか教師値（正解値）が付与されていないケースでも分析モデルの構築・効率的な改善および予測が行えるようになる。 通常、大量の学習データに対して教師値を付与する作業は非常に煩雑になり、大きな労力を要する。対話型モデルの分析機能では、少量の教師値データから初期モデルを生成し、予測精度をより大きく向上させるデータに対して優先的に教師値を付与（ラベル付け）することができるため、同じ労力でより大きな精度向上が見込めるとされる[35]。

目的変数を各公報が審査官引用されたか否かの有無（２値）、説明変数を各公報を分かち書きした単語（TMS の置換語）として審査官引用の有無を予測して結果を評価中である。類似検索と比較して、INFOPRO2016 で発表する予定である。

8 まとめ

機械学習を用いた効率的な特許調査方法をまとめると技術動向調査への機械学習（教師データなし学習）の応用の観点からは従来から良く用いられている書誌事項の統計解析（パテントマップソフト等）と併用することで実務上十分に有用である。さらなる応用として F ターム が付与されている日本特許を教師データとして F ター

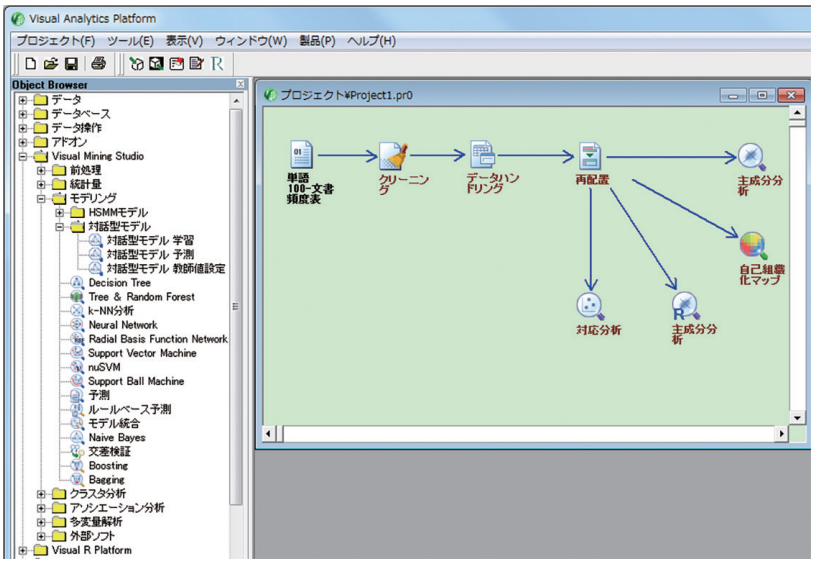

図 11　Visual Analytics Platform

ムが付与されていない中国特許に機械学習で仮想的に付与して活用する等が考えられる。

先行技術調査への応用検討では精度（適合率）と学習コスト上（教師データの準備、学習の手間等）課題がありさらに検討を要する。精度向上からは tf-idf（tf: Term Frequency、単語の出現頻度と idf: Inverse Document Frequency、逆文書頻度の２つの指標にもとづいて計算される）による（コサイン）類似度でなく新規性の観点によく合うように特徴語の重みを機械学習により調整して類似度計算を行う方法。あるいは類似度でなく新規性の観点に適合する評価関数を設計することが考えられる。

9 終わりに

技術動向調査の観点から国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の学術・産業技術俯瞰システム開発プロジェクト[36)、37)]に注目している。成長が見込まれる技術領域（萌芽領域）を特定するのに予測モデルを適用しているのが興味深い。過去において SDI に注目していた新規プロジェクトの特許担当者の立場としては萌芽特許が公開された時点で判れば非常に有用である。

先行技術調査の観点からは将来審査官が拒絶に引用する可能性の高い公報を出願前調査時に見つけることが理想である。教師ありの機械学習は有望そうな感触は得られたが筆者は試行錯誤で機械学習の学習を始めたばかりで今後さらに効率的な調査方法の検討を続ける予定である。

これまで様々なデータベースベンダーから類似検索、概念検索、セマンティックサーチ、スマートサーチ等々いろいろな名前でそれぞれ、類似特許を検索する、概念を検索する、意味を理解して検索する、スマートな検索をする、と受けとれる機能が提供されてきた。実際に検索性能を調べるとベンダーシステムによりさまざまであるが、まだ名前負けしているように思われる。特許データベースのエンドユーザーの立場からは中身の検索技術（アルゴリズム）は問わないが言葉本来の意味で期待に応える検索システムの出現が待ち遠しい。ユーザーにできることは少なくとも名前に惑わされることなくデータ収録、検索性能を正しく理解して目的に応じて適切に使い分けることである。

本報告は 2016 年度の「アジア特許情報研究会」のワーキングの一環として報告するものである。

最後に大変有用な各種ツールを数度に渡り試用させていただき機械学習の初心者である筆者を様々な形でサポートしていただいた NTT データ数理システムの多くの皆様に感謝申し上げます。

参考文献

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

2）知的財産推進計画 2016
http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku20160509.pdf

3）「平成 28 年度 人工知能技術を活用した特許行政事務の高度化・効率化実証的研究事業」の公募結果について
https://www.jpo.go.jp/koubo/koubo/h28_ai_koubo_kekka.htm

4）人工知能学会「What's AI　人工知能研究」
https://www.ai-gakkai.or.jp/whatsai/AIresearch.html

5）機械学習（wikipedia）
https://ja.wikipedia.org/wiki/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92

6）比戸将平ら．データサイエンティスト養成読本 機械学習入門編、技術評論社、2015

7)Willi Richert ら. 実践 機械学習システム、オライリージャパン、2014

8）Sebastian Raschka.Python 機械学習プログラミング 達人データサイエンティストによる理論と実践、インプレス、 2016

9）金 明哲．R によるデータサイエンス - データ解析の基礎から最新手法まで、森北出版、2007

10）金 明哲．テキストデータの統計科学入門、岩波書店、2009

11）R、R 言語、R 環境・・・・・・
http://www1.doshisha.ac.jp/~mjin/R/index.html

12）W eka
https://ja.wikipedia.org/wiki/Weka

13）高橋 佑幸．知りたい分かりたい人の　体験する機械学習、リックテレコム、2015
14）荒木 雅弘　フリーソフトではじめる機械学習入門、2014
15）KH Coder
http://khc.sourceforge.net/
16）樋口耕一．社会調査のための計量テキスト分析、ナカニシヤ出版、2014
17）難波英嗣．人工知能による文書分類、情報の科学と技術 66 巻 6 号（2016）p277-281
18）特許情報分析ツール：Patent Mining eXpress（PMX）
http://www.msi.co.jp/tmstudio/PMXpamphlet.pdf
19）テキストマイニングツール：Text Mining Studio
http://www.msi.co.jp/tmstudio/aboutTMS.html
20）豊田裕貴ら．特許情報のテキストマイニング、ミネルヴァ書房 、2011
21）汎用データマイニングシステム：Visual Mining Studio（VMS）
http://www.msi.co.jp/vmstudio/functions.html
22）アジア特許情報研究会
http://www.geocities.jp/patentsearch2006/asia-research.html
23）Questel 社 Orbit.com
https://www.orbit.com/#WelcomePage
24）情報マネジメント用語辞典：GIGO
http://www.itmedia.co.jp/im/articles/0609/11/news088.html
25）平成 26 年度特許出願技術動向調査報告、人工知能技術
www.jpo.go.jp/shiryou/pdf/gidou-houkoku/26_21.pdf
26）岩本圭介．特許情報テキスト可視化のためのマイニング手法
http://www.japio.or.jp/00yearbook/files/2015book/15_3_05.pdf
27）安藤俊幸．テキストマイニングを用いた効率的な特許調査方法
http://www.japio.or.jp/00yearbook/files/2015book/15_2_12.pdf
28）安藤俊幸ら．精度を重視した効率的な特許調査方法：引用情報と公報の類似度に着目した特許調査方法
https://www.jstage.jst.go.jp/article/infopro/2015/0/2015_47/_article/-char/ja/
29）PATENT EXPLORER
http://www.kibit-platform.com/products/patent-explorer/
30）株式会社 UBIC 機関投資家向けセミナー 2014 年 3 月 6 日
https://www.fronteo.com/ir/ir-data/pdf/20140306_Technology%20briefing%20for%20investors_6Mar2014_Web.pdf
31）統計的テキスト解析（9）～テキストにおける情報量～
https://www1.doshisha.ac.jp/~mjin/R/64/64.html
32）殿川 雅也．これからの特実検索システムの探求 特技懇 no.280.p33
33）条文コード（拒絶理由）
http://www.inpit.go.jp/content/100030260.pdf
34）Text Mining Studio 6.0 新機能
http://www.msi.co.jp/tmstudio/newfeatures_6_0.html
35）VMStudio 8.2 新機能紹介
http://www.msi.co.jp/vmstudio/vmstudio82.html
36）学術・産業技術俯瞰システム開発プロジェクト
http://www.nedo.go.jp/activities/ZZJP_100055.html
37）膨大な論文や特許の引用関係に着目し分析 将来、成長領域となり得る技術を早期に発見
http://business.nikkeibp.co.jp/atclbdt/15/258689/070300001/?ST=print

上記 URL はいずれも 2016 年 8 月 25 日に確認したものである。