機械学習を利用した効率的な特許調査方法：
動向調査と先行技術調査への機械学習の応用

○安藤俊幸[1)]
花王株式会社[1)]
〒131-8501　東京都墨田区文花 2-1-3
Tel: 03-5630-9538　FAX: 03-5630-9712
E-mail: ando.t@kao.co.jp

Effective patent search methods using Machine Learning:
Application of Machine Learning to prior art search and trend survey

ANDO Toshiyuki[1)]
Kao Corporation [1)]
2-1-3, Bunka, Sumida-ku, Tokyo 131-8501 Japan
Phone: +81-3-5630-9538　Fax: +81-3-5630-9712
E-mail: ando.t@kao.co.jp

【発表概要】

機械学習を利用した効率的な特許調査方法を実務ベースに重きを置いて検討した。特に動向調査と先行技術調査への機械学習の応用に関して実務上どこまで使えるかを念頭に検討した。

機械学習を大別すると教師データなし学習と教師データあり学習がある。教師データなし学習の一つの技法としてクラスタリングがある。動向調査の一例として人工知能分野で全体の俯瞰可視化から技術動向を抽出する方向で行った。俯瞰可視化として特許公報間の類似度を用いたクラスタリング検討を行った。

教師データあり学習の先行技術調査への応用として即席麺の直近 10 年を母集団として検討した。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選択して使用した。

動向調査への応用は従来からよく行わ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。先行技術調査への応用検討では網羅性（再現率）は母集団を大きくして対応するとしても精度（適合率）と学習コスト上（教師データの準備、学習の手間等）課題があり詳細に報告する。

【キーワード】

先行技術調査，特許情報解析，機械学習，教師なし学習，教師あり学習，クラスタリング，クラスター分析，可視化

1. はじめに

近年、第3次人工知能ブームにあり囲碁において人工知能が人間のチャンピオンに圧勝、自動車の自動運転など話題に事欠かない状況である。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」[1]においても 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」[2]の公募や中韓文献翻訳・検索システム等の機械翻訳も機械学習の技術がベースになっている。

特許調査の分野においても2015年の特許情報フェアで「人工知能による特許調査・分析業務の効率化と最新事例」をテーマに講演があり会場を変更するほどの大盛況であった。

現状の「人工知能」は機械学習、自然言語処理、情報検索等の様々な情報処理技術から成る複合技術である。機械学習とは、人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとする技術・手法のことである。[3]

2. 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

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ができる手法が理想的である。英語、日本語、中国語で解析可能で日本以外の特許もFタームのような観点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. 方法

データソースの商用データベースとしてグローバルのファミリータイプとしてQuestel 社 Orbit.com、日本特許はNRI サイバーパテントデスク 2、中国特許は日本版 CNIPR を使用した。

特許情報の解析ツールとし下記ツールを使用した。

①Questel 社 Orbit.com の Analysis module

データベースに連動して使用できる専用ツールである。書誌情報の統計解析、テキストマイニング的に抽出した英語コンセプト情報を使用して教師データなしの機械学習を利用したクラスタリング機能、IPC に基づいたクラス分類等が可能である。

②～④は NTT データ数理システムの解析ツールである。

②特許情報分析ツール：Patent Mining eXpress（PMX）[4]

汎用のスクリプト言語 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。

③テキストマイニングツール：Text Mining Studio(TMS)[5]

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」[6]に

ツールと特許情報の解析事例の紹介がある。

④汎用データマイニングシステム: Visual Mining Studio(VMS)[7]

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。

⑤KH Coder[8]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。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。

⑥自作解析ツール [9),10)]

・PatAnalyzer 中国語/日本語解析ツール

・SimCalc1 類似度計算プログラム

⑦R 言語:統計解析 [11)]

⑧Cytoscape:ネットワーク分析 [12)]

4. 結果

4-1. 技術動向調査事例

特許庁の平成 26 年度特許出願技術動向調査報告、人工知能技術 [13)]を参考に Orbit で下記検索式の 1449 ファミリーを Analysis module を使用して解析した。出願数ベースでは 12867 件である。

(G06N)/IPC/CPC AND (US AND JP AND CN)/PN 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。

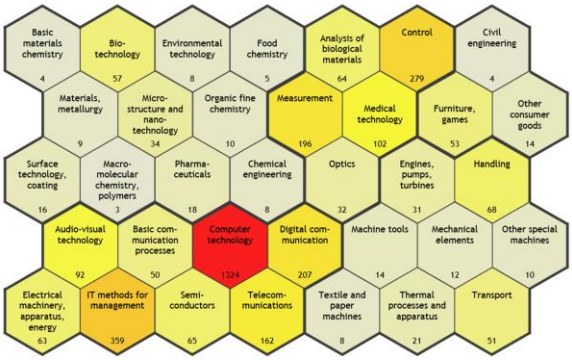

図1. Technology domain チャート

図1. 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。

図2. コンセプトのドーナツチャート

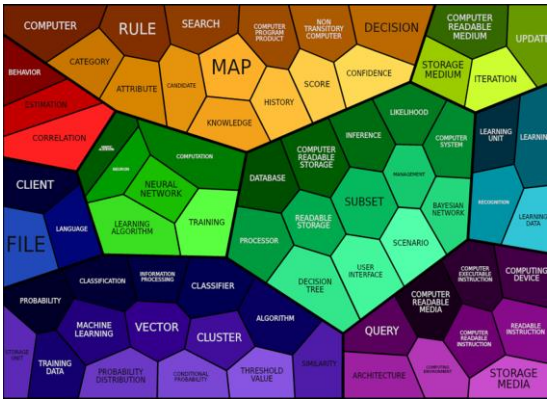

図 3. FoamTree: interactive Voronoi treemap

図3の FoamTree は、JavaScript の高

速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。このような文書のクラスタリング、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。2万件を超える事例[17]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。図2、図3共に内部的には2層になっており上位層（概念）の下に階層（概念）がある。図2は内側が上位で外側が下位で階層関係が明らかである。図2、図3は同じ色でカラーマッピングされて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4 Landscape map は英語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。Landscape map からマウス操作でインタラクティブに任意の領域あるいはクラスターを選択してサブ集合として各種解析処理を行ったり、各々の公報にアクセスできる。約 2 万件のファミリーで約 30 分程度の処理時間を要する[17]。

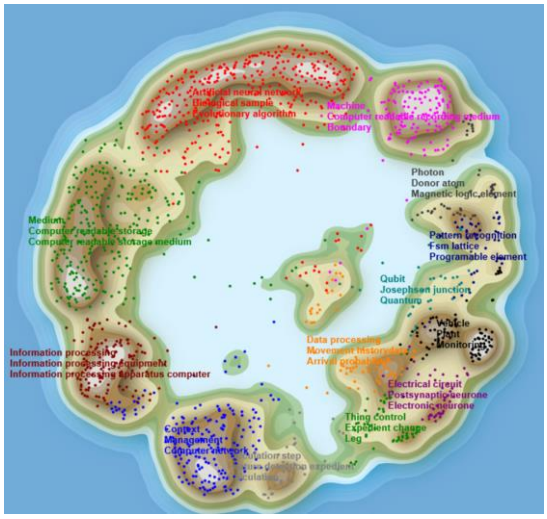

図4. Landscape map

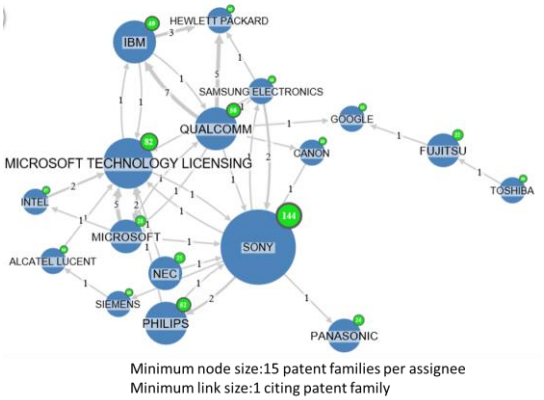

図5. 出願人引用の Node chart

図5は出願人引用をネットワーク表示したものである。最小ノードを 15 ファミリー、最小リンク数 1 ファミリーで主要な出願人間の引用ネットワークを描かせている。

図6は Orbit の検索集合 1449 件（ファミリー）をダウンロードして JP 公報を出願番号ベースで抽出した 1804 件を NRI よりダウンロードして 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・表示したものである。PMX は課題と解決手段の特徴語のクロスマップ、公報間の位置関係を図示するポジショニングマップ機能等を備えている。

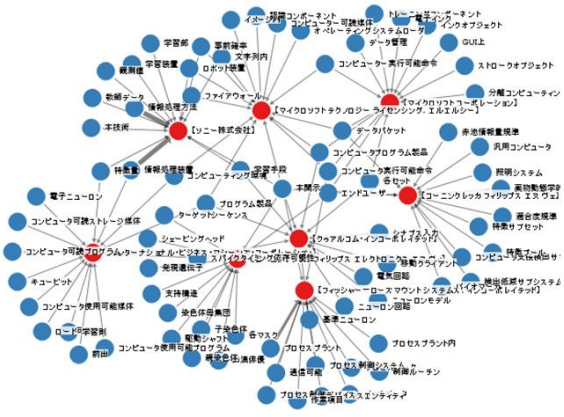

図6. 技術特徴ネットワークグラフ

中心の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。

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。

4-2. 先行技術調査事例

株式会社 FRONTEO(旧 UBIC)の人工知能による特許調査・分析システム PATENT EXPLORER[14]関連情報と「これからの特実検索システムの探求」[15]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③テキストマイニングツール:Text Mining Studio (TMS)と④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(826 件)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絞り込んだ。条文コードだけでは新規性と進歩性を分離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。

VMS は各種の機械学習の手法を使うことができる。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[16]

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。類似検索と比較する予定である。

5. 考察

技術動向調査への機械学習の応用の観点からは 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 タームが付与されていない中国特許に付与して活用する等が考えられる。

先行技術調査の観点からは TF-IDF 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。

6. 結論

動向調査への機械学習の応用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。

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。

7. おわりに

本稿では主に市販のツールで英語、日本語による検討を行った。今後、中国

特許調査への応用検討と類似度計算方法の最適化、新規性の評価関数の構築検討を行い、調査手法をさらに洗練させたい。

「謝辞」
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。
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す。研究会のメンバーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

8. 参考文献

[1] 知的財産推進計画 2016
http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku20160509.pdf accessed 2016.09.14
[2]「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
accessed 2016.09.14
[3] 機械学習（wikipedia）
https://ja.wikipedia.org/wiki/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92 accessed 2016.09.14
[4] 特許情報分析ツール：Patent Mining eXpress（PMX）
http://www.msi.co.jp/tmstudio/PMXpamphlet.pdf accessed 2016.09.14
[5] テキストマイニングツール：Text Mining Studio(TMS)
http://www.msi.co.jp/tmstudio/aboutTMS.html accessed 2016.09.14
[6] 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
[7] 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 accessed 2016.09.14
[8] KH Coder
http://khc.sourceforge.net/
accessed 2016.09.14
[9] 安藤俊幸ら.”精度を重視した効率的な特許調査方法” 第 11 回情報プロフェッショナルシンポジウム
https://www.jstage.jst.go.jp/article/infopro/2015/0/2015_47/_article/-char/ja/ accessed 2016.09.14
[10] 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 accessed 2016.09.14
[11] R 言語
The R Project for Statistical Computing
https://www.r-project.org/ accessed 2016.09.14
[12] Cytoscape 3.4.0
http://www.cytoscape.org/ accessed 2016.09.14
[13] 平成 26 年度特許出願技術動向調査報告、人工知能技術 accessed 2016.09.14
www.jpo.go.jp/shiryou/pdf/gidou-houkoku/26_21.pdf
[14] PATENT EXPLORER
http://www.kibit-platform.com/products/patent-explorer/ accessed 2016.09.14
[15] 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
https://smartcore.jp/tokugikon/uploads/ckfinder/files/gikonshi-latest/280tokusyu03.pdf accessed 2016.09.14
[16] VMStudio 8.2 新機能紹介
http://www.msi.co.jp/vmstudio/vmstudio82.html/ accessed 2016.09.14
[17] 安藤俊幸.Japio YEAR BOOK 2016 機械学習を用いた効率的な特許調査方法