特許分類の自動推定に向けた取り組み
－機械学習による自動分類推定の課題と今後の展開－

Efforts toward automated classification of patent documents

一般財団法人工業所有権協力センター　研究所総括研究員　**小林　英司**

PROFILE　平成25年7月より現職

1　はじめに

一般財団法人工業所有権協力センター（IPCC: Industrial Property Cooperation Center、以下「財団」という。）の研究所では、財団の主たる事業である特許文献の検索事業、分類付与事業の効率化及び高精度化をめざし、独自データ資産を整備するとともに、それらの一層の活用手法を検討している。

特許文献の検索にはIPC、FI、Fターム等の分類を用いるが、分類は技術の発展に追随するために時宜を捉えて改正を行うことが必須であり、分類改正を行った場合には、過去の文献に新たな分類を付与（再解析）する必要がある。そして、再解析すべき対象案件が年々増え続けている中、再解析にはより一層の期間及びコストが必要となってきており、新たな分類体系を用いて検索できるのは、何年も先という現状がある。

そのような状況を踏まえ、財団では分類付与業務の効率化及び高精度化を目的とした、分類の自動推定に関する調査研究を継続的に行い、Japio YEAR BOOKでも紹介させていただいてきた。

本稿では、数年にわたって実施した機械学習技術を取り入れた特許分類の自動推定結果を総括し、今後の自動分類推定に係る研究の方向性について報告する。

2　機械学習による特許分類の自動推定手法

2.1　機械学習による分類推定の概要

財団では、特許分類の自動推定を目指し、機械学習技術の適用を試みた。これは、特許文献に（人手によって）付与されている分類情報を正解データとして機械に自動学習させて分類付与ルールを作成、そして、その分類付与ルールに基づいて、新たな特許文献に対して機械が分類を推定するものである。

ここで、正解データのうち、所定の分類が付与されているものを「正例」、逆に、所定の分類が付与されていないものを「負例」と呼んでいる。

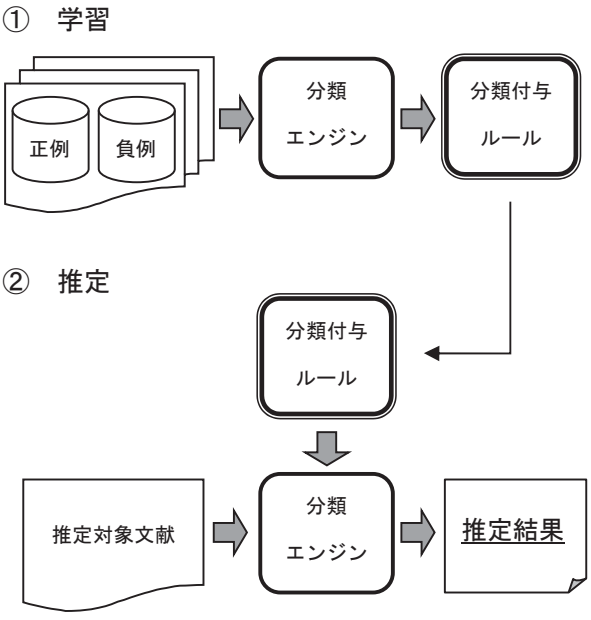

図1　機械学習による自動分類推定システム

2.2　評価手法

上記システムによる分類推定の結果は、システムが推定した分類と正解データとを突き合わせて評価することとした。評価指標は以下のとおり。

・Precision（精度）･･･ 付与すると推定したもののうち、正解分類に存在していた割合。割合が高いと、ノイズが少ないと評価できる。

・Recall（再現率）･･･ 正解分類に存在したもののうち、付与すると推定できた分類の割合。割合が高いと、漏れが少ないと評価できる。

・F 値（Precision と Recall の調和平均）･･･ 下記式で示される、Precision と Recall との総合評価値。

$$F\text{値} = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

2.3　付与根拠データを用いた分類推定結果

機械学習による分類推定には、学習データの内容が重要となる。そこで、財団が有する情報資産である付与根拠データ[1]に着目し、この付与根拠データが含まれる段落全体の文章を学習データとして、Fターム推定を試みた。

表１　付与根拠データを用いたＦターム推定結果

	F 値
テーマＡ（光学系）	0.562
テーマＢ（機械系）	0.541
テーマＣ（化学系）	0.397
テーマＤ（電気系）	0.365

このように、テーマによってＦ値にばらつきがあり、概ね半数以上の推定結果は正しいとされるテーマＡ・Ｂがある一方で、テーマＣ・Ｄにおける推定結果は低いものとなった。これは、学習データの量の差や形態素解析における専門用語の単語抽出の精度等によるものだが、学習データ量の差をどのように埋めるか（付与根拠データが少ない又は無い分野への適用）及び推定精度の向上が課題となった。

1　分類付与者は、所定の分類を付与する際、分類を付与することとなった根拠箇所を明細書等から抽出することがある。付与根拠データは、この抽出した箇所（単語や文書）をテキストとして記録したもの。

2.4　公報全文データを用いた分類推定結果

上記課題を踏まえ、付与根拠データが少ない（又は存在しない）分野への適用を前提に、特許公報全文のテキストデータを学習データとして用いた機械学習による分類推定を実施した。

表２　公報全文データを用いたＦターム推定結果

	F 値
テーマＡ（光学系）	0.537
テーマＢ（機械系）	0.520
テーマＣ（化学系）	0.620
テーマＤ（電気系）	0.399

学習データとして、付与根拠データが含まれる段落全体の文章を利用した場合と、特許公報全文のテキストデータを利用した場合とのＦ値を比較すると、両者に大きな差はなく、特にテーマＣについては大幅に向上する結果となった。付与根拠データが存在しない場合であっても、全文テキストデータを活用することで、分類推定が可能なことが明らかとなった。

2.5　精度向上に向けた取組①

以上のとおり、機械学習による分類推定に一定の有効性が確認できたが、Ｆ値として 0.5 を下回るテーマもあり、実用化には精度向上が課題となる。

そこで、素性を「１」又は「０」で表現し、線形二値分類器である SVM（Support Vector Machine）を用いて境界を計算していた分類エンジンを、TF･IDF（素性出現頻度）法を活用し、「１」又は「０」ではなく素性を重要度（実数値）で表現して境界を計算する分類エンジンに変更することを試みた。

表３　公報全文データを用いたＦターム推定結果

	F 値	
		精度向上効果
テーマＡ（光学系）	0.577	+0.040
テーマＢ（機械系）	0.543	+0.023
テーマＣ（化学系）	0.659	+0.039
テーマＤ（電気系）	0.457	+0.058

表３のとおり、テーマＡ～ＤのいずれにおいてもＦ値が向上しており、TF・IDF 法によって素性に重みをつ

けることが、特許公報の全文テキストデータを用いた機械学習による分類推定の推定精度向上に有効であると言える。

2.6　精度向上に向けた取組②

個々のタームの推定精度は低い場合であっても、そのタームの上位階層のタームの推定精度が高いことが経験則上判明していたことから、さらなる精度向上を目指し、まず上位階層のタームについて機械推定し、付与すべきと推定した場合はさらにその下位階層において機械推定する方式、すなわち、分類の階層構造を利用した分類推定を試みた。

（1）学習フェーズ

階層を利用するため、あるターム及びそのタームの下位階層に位置するタームが付与されている明細書を正例とし、当該Ｆタームと兄弟関係にあるＦタームが付与されている文献を負例とする機械学習を実施した。

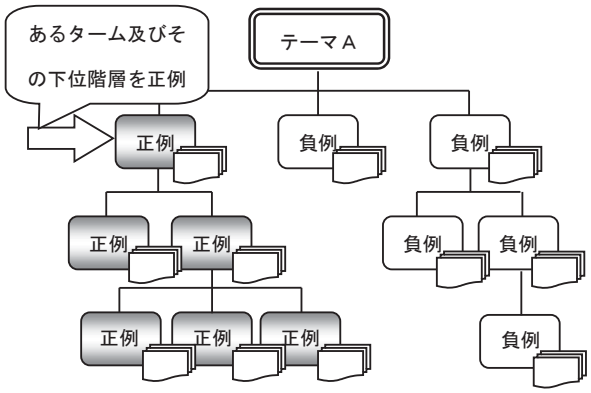

図２　分類階層構造における正例・負例の考え方

（2）推定フェーズ

分類推定にも、階層構造の考え方を取り入れた。

① 最上位の階層に位置するドット０にあるＦタームに対して分類推定を実施し、階層構造を考慮するため、ドット０のターム配下（あるターム及びそのタームの下位層に位置するいずれかのターム）に付与するかどうか推定する。

② ①で付与すると推定した場合、そのタームの下位になるドット１のＦタームについて分類推定を実施する。下位（ドット２）にタームがある場合は○○配下に付与するかどうか推定し、下位タームが無い場合は、そのターム自体を付与するか否かとして推定する。

③ 以降同様に、下位の階層のタームに対して分類推定を実施し、下位Ｆタームが無くなるまで、又は、下位に位置するＦターム全てに対して付与しないと判断するまで繰り返す。

④ 最終的に付与すると推定したＦターム（特定のＦターム又は○○配下）を、機械推定の結果とする。

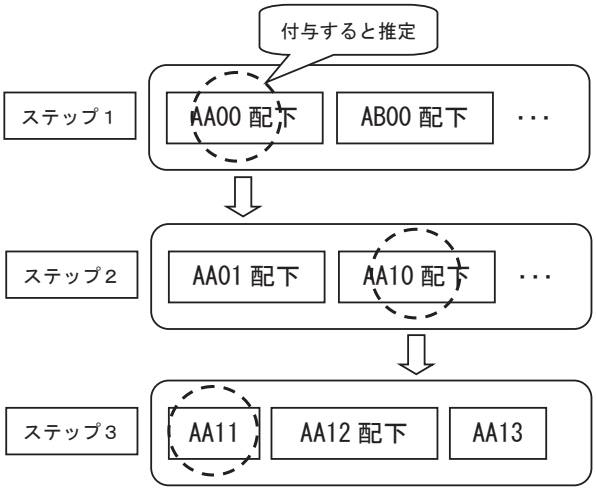

図３　分類階層構造を利用した分類推定の考え方

（3）推定結果

特殊な分類付与ルールを持たない２テーマを選定し、従来の推定と分類階層構造を利用した分類推定を実施した。ただし、推定結果の正解／不正解は、推定したＦタームが正解Ｆタームと一致するか否かで判断するが、後者には「○○配下」という推定結果となる場合があるため、その場合は、○○配下に正解Ｆタームが存在するか否かで推定結果の正解／不正解を判断している。

表４　階層を利用したＦターム推定結果

	Ｆ値	
	階層構造を利用	階層構造を利用しない
テーマＥ（機械系）	0.592	0.587
テーマＦ（化学系）	0.531	0.531

（4）階層構造を利用した分類推定のＦＩへの展開

Ｆタームの分類推定において、階層構造を利用することで精度向上の傾向が見られたことから、Ｆタームと同様、階層構造を有する分類であるＦＩを対象とした分類推定を実施した。

表５　階層を利用したＦＩ推定結果

	Ｆ値	
	階層構造を利用	階層構造を利用しない
テーマＥ（機械系）	0.613	0.587
テーマＦ（化学系）	0.479	0.433

Ｆターム推定で採用した２テーマのＦＩを対象とし、従来の推定と分類階層構造を利用した分類推定を実施したところ、階層構造を利用することによる精度向上の傾向が顕著に現れた。

2.7　総括

以上のとおり、機械学習による特許分類の自動推定に係る研究では、テーマ間の差があり、チューニングよる更なる精度向上の余地はあるものの、分類付与者への支援、すなわち、分類付与者へ付与候補を提示できるレベルまで分類を推定できることが分かった。

ところで、機械学習による分類推定は、十分な正例・負例を用意して機械に学習させ、分類推定する分類エンジンを構築することが前提となるため、特許公報と、その特許公報に付与された分類情報とを一定量確保する必要がある。したがって、機械学習による分類推定は、新設されたＦＩ／Ｆタームに対して短期的に分類を付与する業務には不向きなものとなる。

3　おわりに

特許分類を自動推定する手法として、機械学習技術を適用する調査研究を進めてきたが、テーマ毎のパラメータ調整（チューニング）や、機械が提示した分類付与候補をどのように分類付与者に提示するか（UI 設計）等の、実用化に向けたシステム設計を検討する段階に至っている。

そして、大量の特許文献に対して短期的に分類を付与するニーズが高まっていることを踏まえ、今後は、特許分類の自動推定する手法として、学習データを必要としない新たな手法を中心に、調査研究を進めていく予定である。

参考文献

[1] 笹野秀生，特許分類の自動推定に向けた取り組み－機械学習による自動分類技術の特許文献への適用－，Japio YEAR BOOK 2012，pp.208-211

[2] 小林英司，特許分類の自動推定に向けた取り組み－機械学習による自動分類技術の実用化に向けて－，Japio YEAR BOOK 2013，pp.234-237

[3] 小林英司，特許分類の自動推定に向けた取り組み－特許分類の階層構造を利用した自動推定－，Japio YEAR BOOK 2014，pp.200-203