

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—

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 **笹野 秀生**

PROFILE

平成 23 年 10 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」と表す。）研究所は、財団の主たる事業である検索事業、分類付与事業のさらなる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。

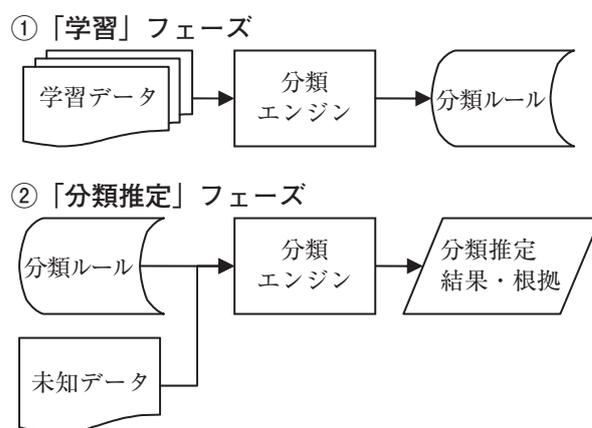
特許文献検索の際には IPC、FI、F ターム等（以下まとめて「分類」という。）を用いるのが常套手段である。分類は技術の発展に追従するために時宜を捉えて改正を行うことが必須であり、分類の改正を行った場合には過去の文献に新たな分類を付与（再解析）する必要がある。

しかしながら、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おり、折角良い分類体系を構築しても、検索で利用できるのは何年も先という現状がある。

そこで、財団では自身の有するデータ資産及びノウハウを活かし、分類付与業務の効率化及び高精度化の可能性を探るため、昨年度から機械学習による自動分類推定に関する調査研究を行っている。本稿では、これまでに行った自動分類推定に関する調査研究について紹介する。

2 自動分類推定の全体構成

自動分類推定は、①「学習」フェーズと、②「分類推定」



【図1】自動分類推定の全体構成

フェーズに分けられ、図1のような全体構成からなる。

学習フェーズでは、学習データを分類エンジンに入力することにより、付与すべき各分類について、一つの分類ルールが生成される。

分類推定フェーズでは、学習フェーズで生成された分類ルール及び未知データとしての特許文献テキストを分類エンジンに入力することにより、自動分類推定結果と、当該付与の根拠箇所とが自動的に出力される。

3 本調査研究の概要

本調査研究は、これまでFタームの自動推定をターゲットとし、上記「2 自動分類推定の全体構成」に沿ったプロトタイプシステムを構築して実験を行っている。以下、その実験で採用したデータ及び方式等を紹介する。

3.1 学習データ

学習データとしては、財団が有する情報資産であるFタームの付与根拠データを活用した。このデータは、特許文献に対して付与されたFタームの付与根拠箇所となる単語や文章をテキスト化したものである。付与者が指摘した根拠箇所は単語だけの場合や、比較的長い文章の場合等がある。実験では、付与根拠データそのものも使用したが、最終的には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採用した。

ここで、特許文献Aの段落1の一部を根拠として、AA01というFタームが付与されていたとすると、特許文献Aの段落1は、AA01の正解データ（正例）となるが、当該段落の記載を根拠としないその他のFタームに対しては不正解データ（負例）となる。段落単位で学習データを設定するため、同じ特許文献A内でも、AA01の正例となる段落と負例となる段落とが生じ得るが、付与根拠データは正例となるべき箇所全てについて作成されているわけではないため、特許文献A内にAA01の正例となる段落がある場合は、同じ特許文献に含まれるその他の段落は、AA01の負例とはしないこととした。

3.2 分類エンジン

分類エンジンは、自然言語処理において広く利用されているSVM(Support Vector Machine)を利用した。SVMは線形二値分類器の一つであり、分類するクラスの数に二つである問題に適用することができる。例えば、未知データにAA01というFタームを付与する問題を考えると、該未知データはAA01を「付与する(正例)」クラス及び「付与しない(負例)」クラスのいずれか一つに分類される。

なお、分類ルール作成及び自動Fターム付与のためには、入力される学習データ及び未知データ(以下まとめて「事例」という。)の中から、各事例を特徴付ける情報(素性(そせい))としての単語を抽出する必要があるが、この抽出作業(形態素解析)には、オープンソースとして利用できる「MeCab」を用いた。

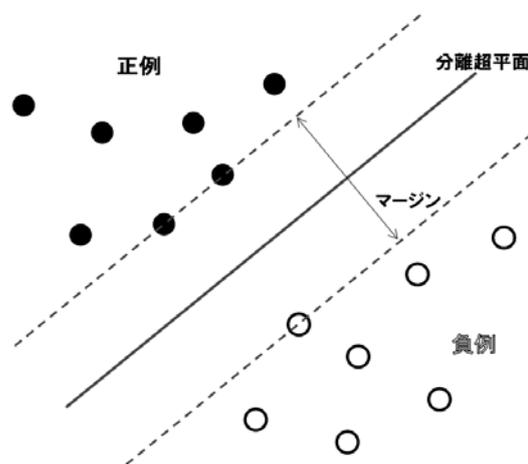
3.3 分類ルール

各事例はそれぞれ、各々に含まれる素性を成分とするベクトル、言い換えると多次元空間内の一つの点 $x^{(1)}$, $x^{(2)}, \dots, x^{(n)}$ で表される(n は事例の数を示す。)。二値分類器において分類ルールを求める作業は、各事例を該多次元空間内で2つのクラスに分離するための超平面(以下「分離超平面」という。)を求めることに他ならない。

この分離超平面は、SVMにおいては分離超平面から最も近い位置にある事例までの距離(マージン)を最大化するように求められる。分離超平面は、多次元のベクトル w (方向ベクトル)と、切片 b とを用いて、 $w \cdot x = b$ と表され、事例を分類するための関数は、次の式で表現される^[1]。

$$f(x^{(n)}) = w \cdot x^{(n)} - b$$

上記式により、事例 $x^{(n)}$ は、 $f(x^{(n)}) > 0$ であれば正例に、 $f(x^{(n)}) < 0$ であれば負例に分類されることになる。図2は二次元空間で考えた場合の分離超平面のイメージであり、マージンが最大になるように分離超平面が選択されている。



【図2】分離超平面のイメージ

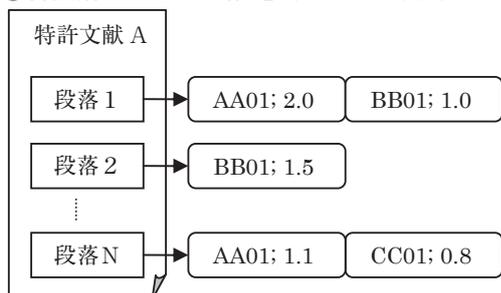
3.4 未知データ

自動Fターム付与を行う未知データは、本来は未分類の特許文献テキストであるべきだが、実験結果(分類推定の精度)を評価するためには、正解情報を有するデータである必要がある。そのため、実験を行うに当たり、上記3.1で説明した複数の学習データの一部を未知データとして採用した。

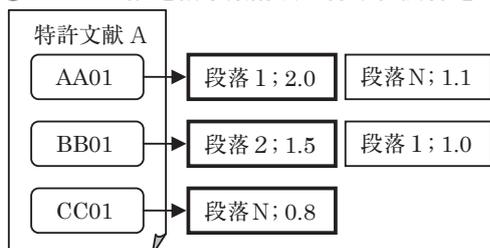
3.5 分類推定結果・根拠

自動分類推定により、未知データ（特許文献テキストの各段落）について、特定のテーマの全てのFタームに関する二値分類がなされるとともに、スコアの計算がなされる。ここでスコアとは未知データの分離超平面からの距離であり、スコアが大きいほど推定の確度が高いものである。図3に分類推定結果・付与根拠結果出力のためのアプローチを示す。

①各段落のFターム推定及びスコア出力



②Fターム推定結果集計及び付与根拠特定



※Fターム毎にスコアの高い段落順に並び替えを実施。

【図3】 分類推定結果・付与根拠出力のアプローチ

推定結果は一つの特許文献の全ての段落について出力されるため、一つの特許文献について集計して推定Fタームを各文献について出力するとともに、スコアに基づいて各Fタームの付与根拠箇所の候補となる段落の優先順位を特定した。

4 分類推定結果の評価指標

分類推定を100%の精度で行うことは不可能と云って良い。図4に挙げるように、推定には「ノイズ」と「漏れ」とが付きものである。この「ノイズ」と「漏れ」と

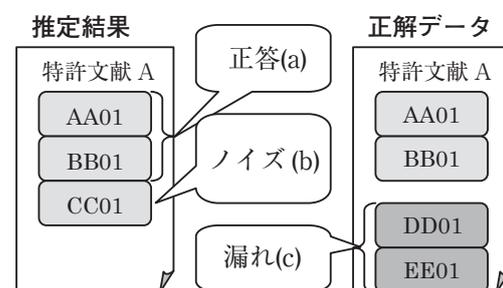
を減らして精度の高い分類推定を行うことが、実用化のキーポイントとなる。この分類推定精度は、次の指標(特にF値)により評価した。

精度：正答(a) ÷ (正答(a) + ノイズ(b))

再現率：正答(a) ÷ (正答(a) + 漏れ(c))

F値：2 × 精度 × 再現率 ÷ (精度 + 再現率)

また、あるテーマ全体の推定精度を評価する場合には、テーマ内の指標の平均値を用いた。



推定結果と正解との比較		正解データ	
		正例	負例
推定結果	正例	(a)・・・2	(b)・・・1
	負例	(c)・・・2	—

※精度：0.67、再現率：0.50、F値：0.57

【図4】 Fターム推定結果評価の一例

5 実験結果

実験では、いくつかのテーマについて、Fターム推定を行い、評価を行った。表1はそのうち4テーマにおけるF値のテーマ内平均値の一覧である。

【表1】 Fターム推定の実験結果

Fタームテーマ	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

実験結果は、テーマによってばらつき（0.365～0.562）が出た。比較的成績の良かった光学系のテーマA及び機械系のテーマBでは、概ね半数以上の推定結果は正しく、正解Fタームの中で半数以上は推定できているといえる。一方、化学系のテーマCや電気系のテーマDなどはF値が0.4を切る結果となったが、その原因としては、学習データが比較的少なかったことや、形態素解析において専門用語の単語抽出が適切にできていなかったこと等が考えられる。

また、個々のFタームに分類するだけでなく、上位階層レベルのFタームに分類する実験も行った。このようにすると分類すべきFターム数が少なくなり、一つのFターム当たりの学習データ数も多くなるため、予測通り精度が上がり、例えば上記テーマAでは最上位Fターム（観点ターム）で0.8以上のF値となり、テーマDにおいても同じく0.7弱のF値であった。

上記実験結果は、テーマ全体の平均値としての精度を見ると、最も良いテーマでもまだ十分な数値とはいえないが、個々のFタームの推定精度は高いものもあり、上位階層レベルへの推定結果なども組み合わせれば、分類付与業務の助けとなり得るものと評価している。

6 おわりに

今回の調査研究により、機械学習を用いた自動分類推定について、一定の有効性が確認された。自動分類推定結果と、当該付与の根拠箇所とを、分類付与実施者に対して提示することにより、効率的かつ網羅的に、付与実施者が付与すべき分類をその根拠箇所とともに検討することができると考えられる。よって、この自動分類推定を高精度に行うことができれば、分類付与業務の大幅な効率化及び高精度化が期待できる。

今後は、分類付与業務の補助システムへの応用を最終的な目標とし、推定精度が低いテーマの底上げや、付与根拠データが無いテーマへの対応等について、調査研究を進めていく予定である。

参考文献

- [1] 高村大也、言語処理のための機械学習入門、2010年8月5日、コロナ社