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―

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、機械学習による自動分類推定に関する調査研究を継続的に行っており、Japio YEAR BOOK 2012 では、付与根拠データを用いた自動分類推定に関する調査研究について紹介した[1]。

本稿では、上記自動分類推定に係る調査研究の結果や課題等を踏まえて実施した、公報のテキストデータを用いた自動分類推定と自動分類推定の精度向上に係る調査研究について紹介する。

1　詳細については、Japio YEAR BOOK 2012 P208-211 を参照されたい。

2 自動分類推定手法と推定精度評価指標

付与根拠データを用いた自動分類推定は、学習データを分類エンジンに入力し、付与すべき各分類について、特定の分類ルールを生成する「学習」フェーズ（①）と、学習フェーズで生成し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を自動的に出力する「分類推定」フェーズ（②）からなり、分類エンジンとして SVM（Support Vector Machine）を、学習データの中から素性[2]を抽出する形態素解析器として MeCab を利用している。

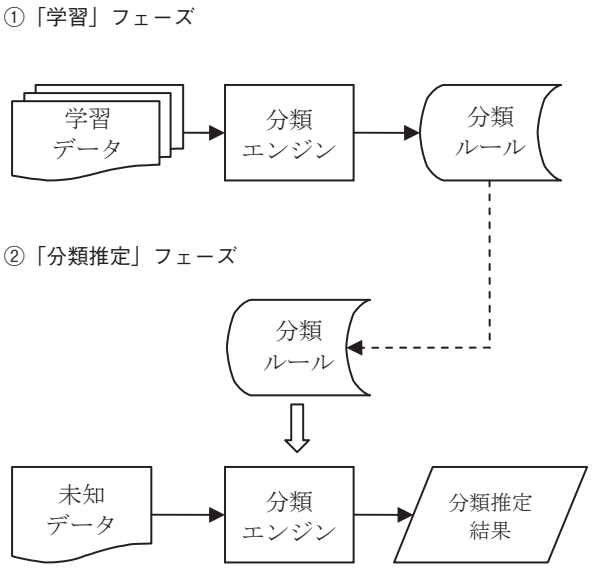


図 1　自動分類推定の全体構成

2　各事例を特徴付ける情報

また、自動分類推定結果の評価指標には、F値[3]を用いた。

3 付与根拠データを用いた自動分類推定の結果と課題

付与根拠データを用いた自動分類推定の結果（技術分野（テーマ）ごとの平均F値）は、0.365～0.562となり、技術分野によるばらつきがあるものの、機械学習を用いた自動分類推定についての有効性を確認することができた。

しかしながら、機械学習による自動分類推定の実用化に向けては、次の課題がある。

① 分類エンジンに入力するための付与根拠データが十分確保されていないテーマコードがあるため、付与根拠データを用いた自動分類推定の適用範囲が限られてしまう。したがって、当該付与根拠データが少ない技術分野（テーマ）への対応を検討する必要がある。

② 推定結果の半数以上が正解Fタームとなったテーマがある一方、F値0.4を切るテーマもあり、全体として、Fターム推定精度の向上を図る必要がある。

4 公報のテキストデータを用いた自動分類推定に係る調査研究

上記課題①への対応として、公開公報（特許請求の範囲及び発明の詳細な説明）から素性を抽出し、学習データとする自動分類推定を行った。

4.1 分類ルールの生成と分類推定

以下の流れで分類ルールを生成し、分類推定を実施した。

① データ加工

公開公報のテキストデータの中から、素性の抽出対象である、特許請求の範囲と発明の詳細な説明の文章を抜き出す（素性抽出テキスト）

② 形態素解析

素性としての単語を抽出するため、①で得られた素性抽出テキストに対して、MeCabによる形態素解析を実施する

③ 学習用データ加工

素性データを、分類エンジン（SVM）に適した形式に変換する

④ 分類ルールの作成

③で得られた学習用データを分類エンジンに入力し、学習をさせ、分類ルールを生成する

⑤ 分類推定

学習用データに用いていない公開公報を未知データとし、この未知データと④で得られた分類ルールを分類エンジンに入力し、分類推定を実施する

4.2 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に、付与根拠データを用いた自動分類推定を実施した4テーマと同じFタームテーマにおけるテーマ内平均値（F値）を示す。

表1 Fターム推定の実験結果

Fタームテーマ	F値	
	公報テキストデータ	付与根拠データ
テーマA（光学系）	0.537	0.562
テーマB（機械系）	0.520	0.541
テーマC（化学系）	0.620	0.397
テーマD（電気系）	0.399	0.365

学習データとして、公報テキストデータを利用した場合と、付与根拠データを利用した場合とのF値を比較すると、テーマA，B，Dについては、ほぼ変わらない結果となり、テーマCについては、大幅に向上する結果となった。

テーマによるばらつきはあるものの、上記結果から、公報のテキストデータを用いた自動分類推定についても、一定の有効性があると言える。

3 F値：2×精度×再現率÷（精度＋再現率）
◇精度： 正答 (a) ÷（正答 (a)+ ノイズ (b)）
◇再現率：正答 (a) ÷（正答 (a)+ 漏れ (c)）

5 自動分類推定の精度向上に係る調査研究

上記課題②への対応として、TF・IDF 法を用いて、F ターム推定の精度向上が可能か調査研究を実施した。

5.1 TF・IDF 法

TF·IDF（素性出現頻度）法は、TF（term frequency）と、IDF（inverse document frequency）、すなわち、ある単語が 1 つの文書に出現する頻度と、ある単語が全文書で出現する頻度を示す 2 つの指標を用いるアルゴリズムのことで、計算式は次の式で表現される。

$$TF_{i,j}=\frac{n_{i,j}}{\Sigma_k n_{k,j}}$$

$$IDF_i=log\ \frac{|D|}{|d:d \ni t_i|}$$

$$TF\text{•}IDF=TF \times IDF$$

$n_{i,j}$ ：単語 i の文書 j における出現回数

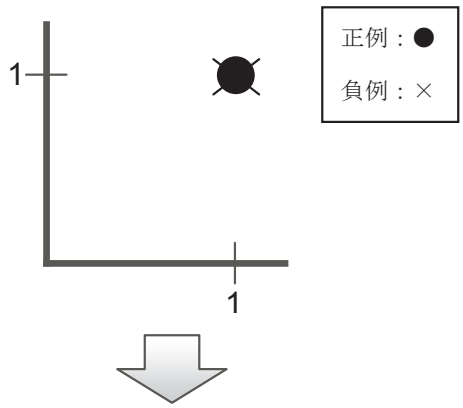
$|D|$ ：テーマ内の総ドキュメント数

$|d:d \ni t_i|$ ：テーマ内の単語 i を含むドキュメント数

この TF・IDF 法では、素性を「1」又は「0」で表すのではなく、実数で表現することから、重要度を反映した正確な分類を可能とする点がメリットとしてあげられる。

また、線形二値分類器である SVM の場合、「1」にある正例、負例から境界を計算していたが、TF・IDF 法を利用することで、実数値にある正例、負例から境界を計算することが可能となる。例えば、正例、負例が同じ値にある場合、境界が計算できないが、TF・IDF 法を利用することで境界が計算でき分類が可能となる（図 2）。

① 正例と負例を分離できない



② 正例と負例を分離することが可能となる

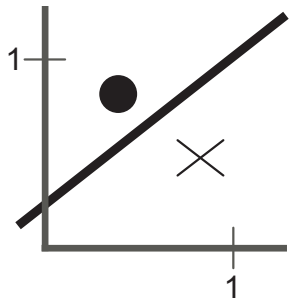


図 2　境界計算イメージ

5.2 実験結果

公報テキストデータを用いた自動分類推定と同じ F タームテーマにおいて、TF・IDF 法で算出した値を素性の重みとして用い、自動分類推定を実施した。その結果を、表 2 に示す。

表 2　公報テキストデータを用いた実験結果

F タームテーマ	F 値（公報テキストデータ）	
	TF・IDF 法を利用	TF・IDF 法を利用しない
テーマ A（光学系）	0.577	0.537
テーマ B（機械系）	0.543	0.520
テーマ C（化学系）	0.659	0.620
テーマ D（電気系）	0.457	0.399

表 2 のとおり、テーマ A ないし D のいずれにおいても F 値が向上する結果となったことから、TF・IDF 法によって素性に重みをつけることが、公報テキストデータを用いた自動分類推定における推定精度の向上に有効で

あることが確認できたと言える。

6 おわりに

本調査研究により、学習データとしての付与根拠データが不十分な場合であっても、公報のテキストデータを用いることで、一定程度の自動分類推定が可能であることが確認できた。

TF・IDF 法でＦ値の改善を図ることはできるものの、Ｆ値は最大で 0.659 であり、機械学習による自動分類推定結果が、ただちに正解分類となるものではないが、人手による分類付与作業時に、参考情報として自動分類推定結果を提示する等、分類付与業務の効率化及び高精度化に向けた活用が想定できる。

今後は、自動分類推定精度の更なる向上や、人手による分類付与作業の支援を前提とした、信頼度の高い分類推定結果の提供方法等、実用化を目指した調査研究をさらに進めていく予定である。