1J1-1

単語出現頻度と機械学習手法を利用した公開特許の課題・手段分類システムの検討

A study of estimation issues and methods from patent documents using machine learning algorithm with term frequency

樽松理樹[*1]
Masaki KUREMATSU

[*1] 岩手県立大学
Iwate Prefectural University

In this paper, I proposed a framework of estimating invention task and means using machine learning algorithm and term frequency in patent documents. It is important to check exists patents before submitting own patent or sealing new products. However, it is take long time to check a lot of patents. In order to support this task, I propose a framework which estimates invention task and means of new patent using machine learning and term frequency in patent journal. First, this framework extracts terms from abstracts of patents which experts identified invention task and means in advance. Secondly, it selects terms based on term frequency and converts labeled patents to training data based on term frequency of selected terms. Thirdly, it makes classifiers using machine learning. Finally, it predicts invention task and means using classifiers. In order to evaluate this system, I did experiments with an expert. In experiments, I used Artificial Neural Network as machine learning algorithm and processed small data set. Experimental result shows it is possible to use this approach to estimate invention task and means from patent journals. However, the accuracy of this approach is lower than the previous approach. The main reason is that learning data, so I will analyze experimental results and consider a new approach to convert learning data from labeled patients to enhance this system.

1. はじめに

特許公報[発明協会 05]は，代表的な知的財産情報である．その情報を有効活用するためには，内容把握，分類が重要である．この作業を支援するシステムがこれまでに提案[藤井 12]されており，その多くはあり，特許公報の請求項や付与されたコードを用いた検索システムである．しかし，実務においては，これらとは異なる分類を用いる場合がある．本研究の研究協力者である企業の知的財産部門に所属する専門家は，その特許が述べている課題と手段で分類している．特許公報が膨大であるため，このような独自の処理に対応するツールが必要となっている．

以上の背景から，これまでに特許公報利用支援の一環として，特許が解決を試みる課題とそれに対する手段を推定する手法[樽松 16]に取り組んでいる．先行手法は，専門家が課題・手段を付与した特許公報から得た語句の出現頻度の類似度に基づき分類することを試みてきた．一定の効果は得られたものの，精度は不十分である．その理由として，語句の出現頻度と分類との関係が十分反映されていないことが考えられる．

本稿では上記の問題点を解決するために，専門家が課題・手段を分類した特許の要約から抽出した語句と課題・手段との関係を，機械学習手法によりモデル化し，それを用いて推定する手法を提案する．また，従来手法との比較を行い，その有用性を評価する．

連絡先：020-0693 岩手県滝沢市巣子 152-52 岩手県立大学 ソフトウェア情報学部，電話 019-694-2582，FAX：019-694-2501，eMail：kure@iwate-pu.ac.jp

2. 語句出現頻度を利用した特許公報からの課題・手段推定システム

2.1 システム概要

本システムは図1に示すように，大きく「分類出現語句情報抽出部」「分類モデル構築部」「分類推定部」からなる．「分類出現語句情報抽出部」では，専門家によって課題と手段ごとに分類された特許公報から，それらの分類を抽出するために有用と思われる分類出現語句情報を抽出する．「分類モデル構築部」では，分類出現語句情報をもとに，機械学習手法を用いて，分類モデルを構築する．「分類推定部」では，分類モデルをもとに課題，手段の候補を推定する．以降で各部分の説明を加える．

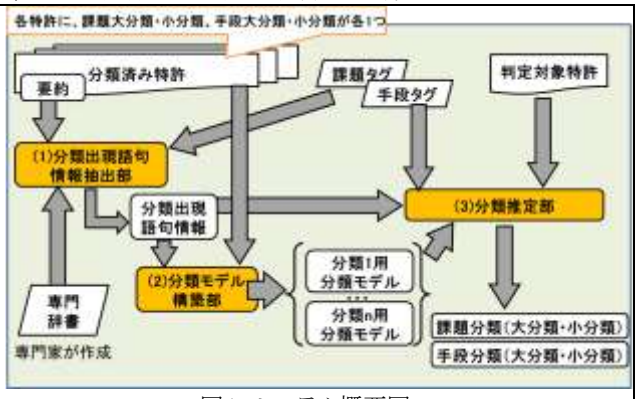

図1 システム概要図

2.2 対象とする特許公報

本システムでは，専門家によって一定の範囲に絞り込まれた特許公報を対象とする．これらの特許公報に対し，専門家は，特許が解決しようとする課題と課題を解決するための手段につ

いて，それぞれ課題の分類を示す課題分類ラベル，手段の分類を示す手段分類ラベルを付与する．課題分類ラベルと手段分類ラベルは，大分類1つと小分類1つから構成される．これらは特許公報に付与されている分類とは異なるものである．

2.3 分類出現語句情報抽出部

専門家に分類付けされた特許公報から，以下の方法で分類出現語句情報を抽出する．

(1) 対象とする文章の抽出…特許公報に含まれる要約文から課題について述べている課題文，手段について述べている手段文にブロックタグを用いて抽出する．

(2) 語句の抽出…課題文，手段文それぞれから，(a)形態素列，(b)カタカナ列，(c)英字列，(d)専門辞書中の代表語のいずれかの方法で 2 文字以上の語句を抽出する．形態素列としては，名詞に着目する．名詞の後に名詞，語尾，形容動詞語幹が連続する場合はそれらをまとめて形態素列として抽出する．ただし，連続する語は 2 語までとする．カタカナ列，英字列はそれぞれ連続する1文字以上のカタカナ，英字の並びである．また，専門辞書とには専門家によって構築された辞書であり，語句と，その語句の概念を示す代表語が与えられている．特許公報中に語句が出現した場合，代表語も抽出する．

(3) 語句の選別…(2)で抽出した語句から以下のいずれかの条件を満たす語句を分類出現語句情報として選別する．

条件1 語句が出現する全文書中において特定の分類の割合が閾値を超える場合．

条件2 ある分野の文章中で，語句が出現する文書の割合が閾値を越える場合．

今回，閾値としては，一様な割合以下であるという仮説が有意水準 5%で棄却できる値を用いた．

以上で求めた語句を分類出現語句情報とする．

2.4 分類モデル構築部

分類情報が付与された特許公報と分類出現語句情報をもとに，機械学習手法を用い，次の手順で分類モデルを構築する．

(1) 分類情報が付与された特許公報に対し，課題推定利用部，手段推定利用をブロックタグに基づき抽出する．

(2) 抽出した部分における，分類出現語句情報の語句の出現回数をカウントする．

(3) 上記の手順で得られた語句毎の出現カウント数を入力信号，分類を教師信号とし，分類ごとに，教師あり機械学習手法を用いて分類モデルを構築する．なお，分類モデルを作成する際には，出現回数をそのまま使うものと，出現の有無を利用するものとを構築する．

結果として，各分類の判定を行う分類モデルを構築する．

2.5 分類推定部

分類推定部では，推定対象の特許を，分類モデル構築部と同様に，課題推定利用部，手段推定利用をブロックタグに基づき抽出する．各部分に対し，分類出現語句情報の語句の出現回数をカウントする．これを分類ごとに構築した分類モデルの入力とする．各分類モデルの出力の集合を推定結果とする．

3. 評価実験

3.1 実験概要

提案手法の有用性を評価するために，3 章で示した考えをもとに JAVA 言語を用いて実装したシステムを用いて，以下の条件のものと実験を行った．実験においては，専門家によって与えられた分類済み特許公報のうち，1998 年から 2008 年までの 283 件から分類出現語句情報を抽出し，2009 年から 2010 年の 59 件の課題分類と手段分類を推定する．専門家が付けた分類を正解とし，第 1 候補として抽出したかにより評価する．

3.2 評価結果

課題には、237 個，手段には 272個の語句が抽出された．

推定結果を表 1 に示す．正答率は，ランダムによる値よりも高いが，先行研究における手法と比較し，どの場合も 10 ポイント以上正答率が低下した．特に，課題大分類のみは 20.3 ポイント低下した．

表 1 ：実験結果

種類	TF	ﾗﾍﾞﾙ数	大小	割合	ﾗﾍﾞﾙ数	大＊	割合
課題	有無	55	5.4%	3/56	12	22.0%	13/59
	回数	55	7.1%	4/56	12	11.9%	7/59
手段	有無	33	14.3%	8/56	8	45.8%	27/59
	回数	33	8.9%	5/56	8	27.1%	16/59

3.3 考察

今回の提案手法では，従来手法を上回る正答率を得ることは出来なかった．要因の一つとしては，学習データ作成方法が挙げられる．今回選別した語句は，一つの分類への出現割合が偏っていることを条件としている．これにより，特定の分野に絞り込むことを試みたが，その語句が出現しない分野の文書への対応が不十分だったと考えられる．そのため，分野の文章を横断する語句の選出が必要である．学習データにおいても分野ごとに数に差があるため，この点の修正が必要である．また，今回は正解として抽出した分類においても，出力値が 0.5 を越えたものは，有無の場合は 3 割，回数の場合は 6 割程度であった．この点から学習に改善の余地があると考えられる．これらのことからデータへの依存度が高いため，その点への改善が不可欠である．

4. おわりに

本稿では，特許公報処理支援を行うために，特許公報で述べられている，解決しようとする課題とその手段の候補を，機械学習手法を用いて推定する手法を提案した．本手法では，専門家により事前に分類された特許公報の要約文における語句の出現情報をもとに機械学習手法で作成したモデルを用いて未分類特許公報の分類を推定する．専門家の協力のもとに行った評価実験においては，正解率は従来手法には及ばなかった．今後は，実験結果の分析に基づく推定方法の改善，処理結果の反映による精度の向上などによる改善を進める．

謝辞

評価実験にご協力いただいた A 氏に感謝の意を表します．また本研究の一部は，科研費・基盤 C（課題番号 15K00154）の助成を受けております．

参考文献

[藤井 12] 藤井敦，谷川英和，岩山真，難波英嗣，山本幹夫，内山将夫：特許情報処理:言語処理的アプローチ，コロナ社（2012）

[発明協会 05]社団法人発明協会：産業財産権標準テキスト 特別編，東京書籍（2005）

[樽松 16] 樽松理樹：語句出現頻度を利用した公開特許からの課題・手段推定システムの検討，人工知能学会全国大会第 28 回（2016）