The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ラフセット理論を用いた特許公報分類支援システムの提案

An Idea of Patent Document Classifier Support System Using Rough Set Theory

樽松理樹[*1]
Masaki KUREMATSUI

[*1] 岩手県立大学ソフトウェア情報学部
Iwate Prefectural University, Faculty of Software and Information Science

In this paper, I proposed a framework of estimating invention task and means using Rough Set Theory. It is important to check exists patents before submitting own patents or sailing 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 decision rules from labeled patent journal using Rough Set Theory and predicts invention task and means from unlabeled patents using them. First, this framework extracts terms from abstracts of patents which experts identified invention task and means in advance. Secondly, it selects terms based on document frequency and makes a Document Term Matrix. Thirdly, it makes decision rules by Rough Set Theory. Finally, it predicts invention task and means using these rules. In order to evaluate this system, I am doing experiments with an expert. I will tell the experimental results and evaluation results about this method on the conference.

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．これまでに提案されてきた特許処理支援システムの多くは，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[藤井 12]である．しかし，実務においては，コードとは異なる分類を用いる場合がある．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，私は企業と協力し，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] [樽松 17]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，ナイーブベイズや ANN を用いて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，ラフセット理論[Zdzislaw 82]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．

2. 提案手法

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「DTM 構築部」「決定ルール抽出部」「分類推定部」からなる．「DTM 構築部」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語句を選択，それをもとに文書語句行列（以後，DTM）を構築する．「決定ルール抽出部」では，ラフセット理論に基づき，構築した DTM から決定ルールを抽出する．「分類推定部」では，決定ルールをもとに新たな特許の課題，手段の候補を推定する．以降で各部分の説明を加える．

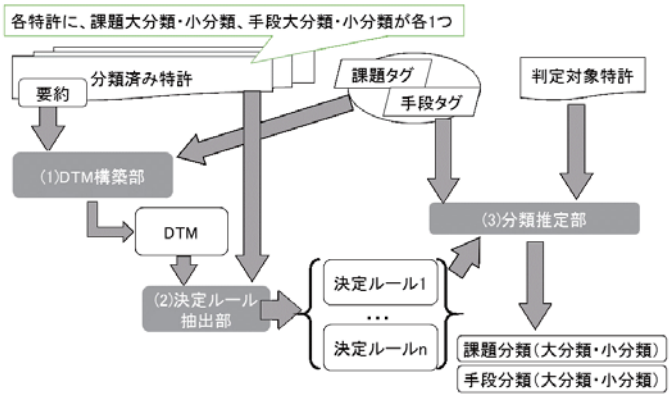

図1 システム概要

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（対象）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つから構成されている．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 DTM 構築部

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を取り出す．要約文は，【課題】文 *1*, ･･･, 文 *n*【解決手段】文 *n+1*, ･･･, 文 *m* または【目的】文 *1*, ･･･, 文 *n*【構成】文 *n+1*, ･･･, 文 *m* ような構造をとる．このうち，文 *1*, ･･･, 文 *n* を課題について述べている課題文，文 *n+1*, ･･･, 文 *m* を手段について述べている手段文として抽出する．

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)英字列いずれかの方法で語句を抽出する. 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. カタカナ列, 英字列はそれぞれ連続する 2 文字以上の並びとする.

(3) 語句の抽出…(2)で抽出した語句から以下の条件を満たす語句を抽出する.

条件 特定の分類の特許に対する DF 値が閾値 θ 以上.

本システムの目的が分類推定であるため, 特定の分類に出現が偏っている語句を抽出することで, 分類の絞り込みを図れると考え, 本条件を利用する. TF 値については, 対象とする文書が短いことを考慮し, 今回は利用しない.

(4) DTM の構築…(3)で抽出した語句を元に DTM を構築する. DTM の各要素は出現の有無を示す 2 値からなる.

2.4 決定ルール抽出部

前述までの処理で構築した DTM に対し, 次に示すラフセット理論の考えに基づき, 決定ルールを抽出する.

(1) 決定表の作成・・・DTM を条件属性集合, 分類を決定属性集合とし, 決定表を作成する. ここで決定属性とは回帰分析における目標変数, 条件属性とは説明変数にそれぞれ相当する. また決定表の各行が, 各特許に対応する.

(2) 決定ルールの作成・・・決定属性の 1 つである分類 c の決定ルールを次の方法で構築する.

(2a) 下近似集合の抽出・・・決定表より, 決定属性が分類 c のうち, 条件属性の値が, 他分類の特許の条件属性の値と一つでも異なる異なる特許のみを抽出する. これを下近似集合と呼ぶ.

(2b) 決定行列の作成・・・抽出した分類 c の下近似集合に含まれる特許 i を行, 分類 c 以外の特許 j を列とし, その交点に特許 i の決定属性の値のうち, 特許 j と異なる部分を記載した決定行列を作成する.

(2c) 決定ルールの作成・・・決定行列の各行に対し, 決定行列の交点の論理積からなるルールを作成する. このとき, 交点内の要素については, 論理和と見なす. 生成された各ルールの論理和をもとめ, 各要素の包含関係から, 要素をまとめていく. 最終的に残った要素を条件部, 決定属性を結論部とする決定ルールを抽出する. また, 抽出したルールを満たす特許を求め, その中で分類 c に属する特許の割合を, 同ルールの重さとする.

(3) すべての分類に対し, 決定ルールを取り出した場合は終了する. また構築していない分類がれば, (2)を繰り返す.

以上の過程で求めた決定ルールを用い, 分類を推定する.

2.5 分類推定部

分類推定部では, 推定対象の特許の要約文から, DTM 構築部と同様に, 課題文, 手段文を抽出する. 各文に対し, DTM 中の語句の出現の有無を確認する. 得られた結果に対し, 決定ルールを次のように適用し, 分類を推定する.

(1) 前件部をすべて満たす決定ルールを求める. 該当するルールがある場合, そのルールの重さの和を分類の評価値とし, 評価値の降順で並べたリストを推定結果とする.

(2) 前件部が一致するルールが無い場合は, 各分類の評価値として, 決定ルールの成立した条件部の割合と重さの積和を求める. 得た評価値の降順で並べたリストを推定結果とする.

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, 実験をおこなっている.

実験においては, 専門家から提供をうけた分類済み特許公報を, 1998 年以前の特許 297 件(Data-1), 1998 年から 2008 年まで特許 283 件(Data-2), 2009 年から 2010 年の特許 59 件(Data-3)に分割し, Data-1 から抽出した決定ルールを用いて Data-2 の分類を, Data-2 から抽出した決定ルールを用いて Data-3 の分類をそれぞれ推定する. それぞれ, 事前に付けられた分類を正解とし, 抽出したリストにおける順位との比較により評価する. 現時点では評価中であり, 発表時に結果を示す.

3.2 考察

評価実験は実施中であるが, 予備実験においては, 今回の提案手法では, 従来手法より良い正答率を挙げている. その理由としては, 従来手法では扱えていなかった「語 x を含まない」という否定が決定ルールに反映できていることが挙げられる. また, 従来手法に比べ, 決定ルールは可読性が高いことから, 専門家による検証も進めやすいものと考えられる.

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, ラフセット理論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ラフセット理論で抽出した決定ルールを用いて未分類特許公報の分類を推定する. 現在, 専門家の協力のもと評価実験を行っており, その結果に基づく評価, 抽出されたルールの検証, 評価結果の分析に基づく推定方法の改善などが今後の課題である.

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 29 回(2016)

[樽松 17] 樽松理樹: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, 人工知能学会全国大会第 30 回(2017)

[Zdzislaw 82] Pawlak, Zdzisław :Rough sets, International Journal of Parallel Programming, Vol. 11, No. 5, pp.341–356, Springer, Heidelberg (1982)