様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１９　日現在

機関番号：　２１２０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００１５４

研究課題名（和文）ラフセット理論を活用した特許実務支援システムの構築

研究課題名（英文）A Study of a Patent Management Support System with Rough Set Theory

研究代表者
榑松　理樹（Kurematsu, Masaki）

岩手県立大学・ソフトウェア情報学部・准教授

研究者番号：００３０５２８６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，特許公報実務支援システムとして，(1)特許公報の内容把握支援機能および（2）特許関連MAP作成支援機能を統合したプロトタイプシステムを構築した．本システムは，はじめに，専門家が分類付けした特許の要約から，特定の分類に出現する傾向がある語句を抽出する．次に，各特許におけるそれらの語句の出現と分類の対応から，ラフセット理論に基づき決定ルールを構築する．このルールを用い，未分類の特許の分類を行う．専門家と協力して行った評価実験においては，正解の分類を上位25%以内に，最低で49.7%，最高で76.6%を抽出した．今後の課題としては，抽出する語句の影響をいかに抑えるかが，あげられる．

研究成果の概要（英文）：It is important to research exists patents before submitting own patents or saleing new products. However, it is take long time to check a lot of patents. In this research, I propose a framework in order to this task. This framework estimates decision rules from labeled patent journals using Rough Set Theory and estimates a category from unlabeled patents using these rules.
At first, this framework extracts terms from abstracts of labeled patents in advance. Next, it selects terms based on document frequency and makes a Document Term Matrix. After that, it makes decision rules by Rough Set Theory. Finally, it estimates a category using these rules.
In order to evaluate this approach, I did experiments with an expert. In this experiment, this system could estimate correct categories from most of patents. However, the performance of this method is similar as exists methods. In order to enhance this system, I should to enhance term selection and evaluate new idea.

研究分野：　知能情報学

キーワード：　自然言語処理　ラフセット理論　文書分類　特許処理

１．研究開始当初の背景

代表的な知的財産情報である特許公報を活用する重要なタスクとして，内容把握，分類，情報蓄積等がある．しかし「内容把握が困難」「個人間での観点の違いにより結果が多様化」「結果等の多様化により蓄積情報共有が困難」等の問題が生じている．特許公報活用の有効性，効率性を向上させるためには，このような問題を解決する必要がある．

この問題に対し，これまでにコンピュータによる支援方法が提案されてきた．しかし，その多くは，特許庁電子図書館(IPDL)サービスに代表されるような検索システムや類似文書検索システムである．これらの多くは，キーワードに着目し，表層情報レベルでの処理を行っている．近年，表層情報に加え，概念辞書やオントロジーなどを用いた処理も試みられている．しかし，検索結果に誤った特許が含まれるなど検索精度に問題が残っているのが現状である．また，これらのシステムは，公開特許のフロントページを対象とした特許検索が主であり，内容把握や分類等の作業は依然として人手で行うことが多い．特許公報活用の有効性や効率性を向上させるためにも，内容把握や分類，情報蓄積等の文書処理支援手法を確立することが求められている．

研究代表者も平成 12 年より，テキストデータを対象とし，キーワード検索や文書検索の研究に着手している．これまでに，「共起関係による手法」「文中の係り受けに関する距離による手法」「語ベクトルの類似度による手法」の手法について提案・改善を行ってきた．これらの手法に対して，テストコレクションを用いて評価した結果，一部の手法で若干の向上が見られたが，全体として正答率は低く，改善の余地が残っている．また平成 19 年度から 21 年度にかけて，基盤研究(C)として「新聞記事を利用した対話型訴訟相談システム」を実施した．その過程において，特定事件の記事検索を実現するために新聞の文書構造や助詞に着目した手法を提案した．従来手法よりも計算量は増えるものの，目標とする記事発見の精度向上が見受けられた．これらの研究を通し，文書検索や処理においては従来のキーワードのみでなく，「文書構造」や「助詞」に着目することが有用であるとの着想を得た．

これらの成果を元に，特許文書に関する研究を平成 23 年度より進めている．平成 23 年度では滝沢村産学共同研究事業補助金を受け，実務者と協力して研究を行った．この研究では特許公報の文書構造は比較的処理しやすいという結論を得るとともに，内容把握で着目する点は「課題」と「手段」であること，「従来技術」は特許の履歴をたどる上で重要であること，企業名も重要な情報となることなどの知見を得た．この成果を踏まえ，平成 24 年度から平成 26 年度にかけ，基盤研究(C)として「文書構造レベルの統計モデルを用いた特許公報管理支援システムの構築」を進めている．当初，クラメールの連関係数の援用手法を考案，実験を企業と協力し実施してきたが，結果として，既存特許に対し，独自の分類が活用されていること，「課題」「手段」単位で処理を行うこと，既存技術である *Cos* 類似度の有用性が高いこと等が明らかになった．また分類と特許内の語句表現との関係が強いことから，分類済みの特許公報から，ナイーブベイズ等の機械学習手法を用いることでそれらの分類を示す語句表現が獲得でき，特許公報の処理効率向上ができるとの着想を得た．

２．研究の目的

本研究では，現在，多大な労力と時間が必要としている特許実務の負荷軽減とともに，特許公報活用の利便性を向上させることを目的に，(1)特許公報で述べられている課題と手段の分類を抽出する内容把握支援機能，(2)抽出した課題と手段の分類に基づく特許関連 MAP 作成支援機能を持つ特許実務支援システムの開発に取り組む．(1)では，専門家が分類分けした公開特許から，ラフセット理論を援用し，語句出現頻度に基づく重みを持つ分類抽出ルールを構築する．このルールと特許の文書構造を考慮した文書類似度を基に，課題と手段の分類を抽出する．(2)では，(1)で得た分類をもとに，分類の割合を時系列で示した動向 MAP，類似度に基づき特許を平面上に配置した類似度 MAP の作成を行う．さらに専門家とともに本システムの有用性を評価する．

３．研究の方法

２で述べたことを踏まえ，本提案研究課題として，特許公報実務支援システムの構築を目指す．本システムの主な機能として，(1)特許公報で述べられている課題と手段の分類を抽出する内容把握支援機能，(2)抽出した課題と手段の分類をもととした特許関連 MAP 作成支援機能がある．以下，各機能について説明する．

(1) 特許公報の内容把握支援機能：本機能は，特許公報の内容として，その特許で述べている「課題」「手段」の分類を抽出する．ここで分類とは，課題や手段の内容をまとめたものであり，大分類と小分類から構成される．特許公報に与えられている IPC や FI といった特許分類とは異なり，特許の実務に携わる専門家が特許の内容を把握するために独自の分類として用いている．本研究では，これらの分類を次に示す方法で抽出する．

専門家がこれらの分類を付与した特許公報から，分類ごとに頻出語句を取り出し，語句集合を生成する．次に語句集合に対しラフセット理論を適用し，分類ごとの出現語句に基づく分類抽出ルールを作成する．これらルールには出現頻度とラフセットの精度に基づく重みを与える．これらルールをナイーブ

ベイズで用いられる手法を援用し，特許公報と照合することで，各分類の類似度を求める．また，これとは別に，特許の文書構造を示すブロックタグに着目し，「課題」「手段」ごとに従来の文書ベクトル法をもとに特許公報ごとの類似度を求め，それをもとに分類ごとの類似度を求める．最後にこれらの類似度を統合し，各分類の類似度を求める．また，結果に対し，ユーザ（専門家）が評価し，それをルールに反映させることで，ルールの精度向上を図る．

(2) 特許関連 MAP 作成支援機能：本機能は，内容把握支援機能で得た「課題」「手段」の分類をもとに特許間の関連，分類の動向把握，可視化を試みる．具体的には，「課題」「手段」について時系列ごとに分類の割合を並べた動向 MAP，特定の特許を中心に「課題」「手段」の分類ごとの類似度をもとに特許を配置した類似度 MAP を作成する．これらの MAP を示すことにより，技術動向，特許間の関連性の把握の効率化を図る．

本研究では上記の機能の設計，開発を行い，実務者を交えた実証実験によって，その有用性を評価する．

４．研究成果

(1)概要

本研究では，特許公報実務支援システムの主たる機能である (1)特許公報の内容把握支援機能を中心に研究を行った．本システムは，専門家が課題・手段を分類した特許の要約から，ラフセット理論に基づく決定ルールを構築，それを利用して新規特許の課題および手段の分類を行う．本システムは，大きく「DTM（文書語彙行列）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 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．本システムの概観を図１に示す．

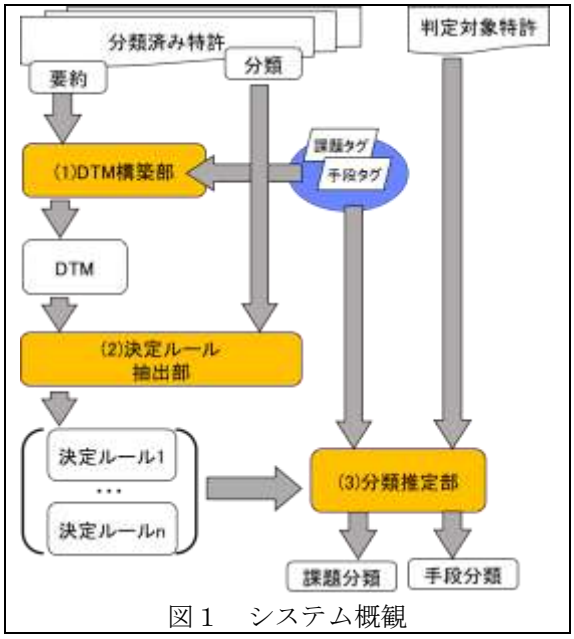

図１　システム概観

(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および課題を解決するための手段について，それぞれの分類を示す課題分類ラベル，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独自のものである．特許公報に付与されている分類は，専門家による判断と異なる点，請求項に基づく点などの理由から，本研究では利用しない．

(3)DTM 構築部

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

（Step1-1）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”の構造をとる．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 を手段段について述べている手段文として抽出する．

（Step 1-2）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)英字列に合致する語句を抽出する．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．

（Step 1-3）語句の抽出

(Step 1-2)で抽出した語句から 2 回以上出現し，分類の出現数が 2 つ以下の語句を抽出する．これは，特定の分類に出現する語句に着目するためである．

（Step 1-4）DTM の構築

(Step 1-3)で抽出した語句を元に DTM を構築する．DTM の各要素は，各特許公報における(Step 1-3)で抽出した語句の出現の有無を示す値からなる．

(4)決定ルール抽出部

決定ルール抽出部では，前述までの処理で構築した DTM と各文書の分類から，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．

（Step 2-1）決定表の作成

DTM（各語句）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．

（Step 2-2）決定ルールの作成

決定属性の各分類に対し，次の方法で決定ルールを構築する．

（Step 2-2a）下近似集合の抽出

決定属性が分類 c のうち,条件属性の値が,他の分類の特許には出現しない特許のみを抽出する.これを分類 c の下近似集合と呼ぶ.

（Step 2-2b）下近似決定行列の作成

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 の条件属性のうち，特許 j と異なる部分を記載した下近似決定行列を作成する．

（Step 2-2c）下近似決定ルールの作成

下近似決定行列の各行に対し，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下近似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．

（Step 2-2d）上近似集合の抽出

決定属性が分類 c の特許および，条件属性の値が分類 c のいずれかの特許と一致する他分類の特許を抽出する．これを分類 c の上近似集合と呼ぶ.

（Step 2-2e）上近似決定行列の作成

抽出した分類 c の上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 の条件属性のうち，特許 j と異なる部分を記載した上近似決定行列を作成する．

（Step 2-2f）上近似決定ルールの作成

上近似決定行列の各行に対し，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上近似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．

以上の工程を，すべての分類に対し，決定ルールを取り出すまで繰り返し行う．

上記の結果，作成した決定ルールを用い，分類を推定する．

(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 構築時と同様に，課題文，手段文を抽出する．各文に対し，ルール作成に利用した 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，下近似・上近似それぞれの決定ルールを次の方法で適用し，分類を推定する．

（Step 3-1）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以下の計算式で統合する．

$E(r1, r2)=E(r1)+E(r2)-E(r1)\times E(r2)$

$E(x,y)$は，x,y の評価値，$r1$ および $r2$ はルールを示す.

評価値を降順にソートした結果を，推定結果とする．

(6)評価実験

以上の提案手法の有用性を評価するために，前述の考えをもとに JAVA 言語を用いて実装したシステムを用いて，実験をおこなった．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 639 件を， 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 および Data-3 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．

また，ナイーブベイズを基本とする方法での推定結果と比較した．ナイーブベイズを基本とする手法では，分類を推定する特許に出現する語句をもとに，以下の計算式で分類ごとの評価値を計算する．

$E(c)=\Sigma w(c,t)$

ここで，c は分類，t は特許中に出現する語句,$w(c,t)$は，語句 t の分類 c に対する重み，$E(c)$は分類 c に対する評価値を示す．また,$w(c,t)$は以下の 3 つの値を利用する．

$w_1(c,t)=\log(1+tf(c,t)/tf(c))$

$w_2(c,t)=\log(1+df(c,t)/df(c))$

$w_3(c,t)=\log(1+cf(t)/分類の総数)$

ここで,$tf(x,y)$，$df(x,y)$はそれぞれ，x,y が出現した回数，文書数を示す．また，$cf(t)$は，t が出現した分類数を示す.

実験の結果を表 1 に示す．なお表の値は，3 パターンの平均値である．ここで，w_1からw_3は比較として行った手法である．

表 1 ：実験結果（平均値）

手法	課題（12 分類）		手段（8 分類）	
	順位	上位 25%以内	順位	上位 25%以内
下近似	4.4	49.7(%)	2.1	76.6(%)
上近似	4.6	45.5(%)	3.0	47.5(%)
w_1	3.5	62.3(%)	2.4	65.5(%)
w_2	3.5	62.3(%)	2.3	65.8(%)
w_3	4.2	52.8(%)	2.8	60.8(%)

評価実験の結果は，下近似決定ルールでは，正解を上位に推定できた．これは，語句がないが有用に働いたためと考えられる．ただし，比較として行った TF や DF を用いたナイーブベイズ手法との大きな差異は得られなかった．一方，上近似決定ルールは，他のものよ

り悪い結果となった．これは上近似の性質を考慮すれば，予想されたことである．

今後の課題としては，ルールの精査，作成・利用法の検討，語句抽出方法の検討，再検証が挙げられる．ルールの精査においては，冗長・包含関係のあるルールの発見，専門家によるチェック，作成・利用方法の検討については，上近似と下近似の使い分け，重みの更新方法の検討，語句抽出方法の検討については，データによる偏りへの対応，辞書の活用／機械翻訳の活用，再検証では，同一データでの実験，新聞記事など特許とは異なる分野での実験が挙げられる．

また，(2) 特許関連 MAP 作成支援機能については，得られた結果をもとに，「課題」「手段」について時系列ごとに分類の割合を可視化する動向 MAP，特定の特許を中心に「課題」「手段」の分類ごとの類似度をもとに特許を配置した類似度 MAP を作成中である．これを完成させることも今後の課題である．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　7　件）

① 榑松理樹，ブロック単位の語句の出現頻度に基づく特許課題・手段推定システム，2015 年度人工知能学会（第 29 回），2015
② 榑松理樹，ブロック単位の語句の出現頻度に基づく特許分類支援システムの提案，第 14 回 FIT 情報科学技術フォーラム，2015
③ 榑松理樹，A Framework for a Decision Tree Learning Algorithm with Rough Set Theory，The 14th International Conference on Intelligent Software Methodologies, Tools and Techniques，2015
④ 榑松理樹，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，2016 年度人工知能学会全国大会，2016
⑤ 榑松理樹，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，2017 年度 人工知能学会全国大会（第 31 回），2017
⑥ 榑松理樹，ニューラルネットワークを用いた特許分類システムの提案，第 33 回ファジィシステムシンポジウム，2017
⑦ 榑松理樹，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，2018 年度人工知能学会全国大会（第 32 回），2018

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕

ホームページ等

６．研究組織

(1)研究代表者

榑松　理樹（KUREMATSU, Masaki）
岩手県立大学・ソフトウェア情報学部・准教授
研究者番号：00305286

(2)研究分担者　なし

(3)連携研究者　なし

(4)研究協力者　なし