様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２３ 日現在

機関番号：２１２０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１２１

研究課題名（和文）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

研究課題名（英文）Estimating Invention Task and Means from Patent Journals Based on the Document Similarity with the Patent Journal Structures

研究代表者

樗松　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）　　2,500,000円

研究成果の概要（和文）：権利調査などにおける特許公報処理支援を行うために，特許が解決しようとする課題とその手段の候補を推定する手法を提案した．本手法では，大分類と小分類の組み合わせから表現された課題分類と手段分類を，専門家が事前に行った課題分類・手段分類の抽出結果をもとに推定する．専門家の協力のもとに行った評価実験においては，課題分類・手段分類の組については，10位以内に正答が含まれる割合は最大で，大分類小分類両方の場合は約17%，大分類のみの場合は約49%であった．今後の課題として，より大規模なデータでの実証実験，語句の切り出し方やブロックタグの利用などによる各種情報抽出方法の検証と改善などがあげられる．

研究成果の概要（英文）：In order to support checking exists patents, I proposed a framework of estimating invention task and means from patent journals based on the document similarity between it and each exist patent given invention task and means by experts. The framework divides a patent into 4 parts based on the patent journal structure initially. It gets the similarity of each part and integrates them and shows invention task and means ranked in descending order by the similarity in the difference 4 orders. I did an experiment with an expert and small data set. In this research, invention task and means have sub categories. I evaluate this system from the viewpoint of the rank of invention task and means given by experts. 17% of the collect invention task and means with sub categories had entered in the top 10 and 49% of them without sub categories had entered in the top 10. This experimental result shows that is it possible to use my idea to estimate invention task and means from patent journals.

研究分野：知能情報学

キーワード：情報検索　特許情報　文書類似度　文書分類

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に対し、内容把握、分類、情報蓄積等を行うことは重要なタスクである。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や分類が多様化する」「把握結果等の多様化による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。

これらの問題に対し、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。これらのシステムの多くは、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、これらのシステムでは特許検索が主であり、内容把握や分類など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、情報蓄積などの文書処理支援手法を確立することが依然として求められている。

２．研究の目的

１章で述べたことを踏まえ、本提案研究課題として、特許公報処理を効率良く支援するシステムを提案する。本システムの主な機能として、(1)特許公報の内容把握支援機能および (2)特許関連 MAP 作成支援機能の実現を目指す。またこれらの結果を活用した(3)特許公報内容に着目した検索機能の実現も目指す。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

特許公報の内容把握を行う上では、「技術課題」「技術課題に対する解決手段」「解決手段に用いた機能・部品」の把握が重要となる。これを行うために、事前に専門家が評価した特許公報を基に、①特許の文書構造に基づく重要個所の抽出、②重要個所からの単語や助詞、フレーズの抽出、③専門家による内容把握および国際特許分類とフレーズ類の出現傾向に基づく統計モデルの構築を行う。また、統計モデルを構築する際には、カタカナに対しては音相での比較も行う。この統計モデルを活用することで、新たな特許公報からの重要個所のフレーズを抽出し、提示することで内容把握支援を行う。また、推定結果に対するユーザの評価を用い、統計モデルの改善を行う枠組みの構築も行う。

(2)特許関連 MAP 作成支援機能

「従来技術」など重要度の低い部分に対し、内容把握支援と同様の処理を行う。この情報と内容把握支援で得た情報とをインデックスとして利用することにより、特許関連 MAP を作成する。基本的には統計モデルに基づき、類似度を算出する。この類似度を基に、時系列や出願者などの書誌情報など対話型インタフェースを通し、ユーザが与える次元に合わせて特許公報を配置することで、特許関連 MAP を作成する。

(3)特許公報内容に着目した検索機能

(2)で作成する統計モデルをインデックスとして利用し、ユーザが入力する「技術課題」「解決手段」に関連する特許公報の検索を行う。さらに(2)で構築する MAP 上に配置することにより、その位置づけを示す機能の実現を試みる。

３．研究の方法

３．１　初年度の内容

研究初年度の平成 24 年度においては、主に以下のことを行う。

①特許文書処理に関する知識の収集

平成 23 年度に共同研究を実施している企業を対象に、特許文書処理に関する知識の収集を進める。具体的には、処理時に注目する文書の構造、処理の流れ、重要と判断するポイントなどである。平成 23 年度に収集している知識にこれらの知識を加えた知識を、以降の処理において活用する。なお現在の共同研究は平成 23 年度の単年度であることから、これらの知識提供に対しては、謝金を支払う。また上記と並行し、特許処理に関する動向調査を行う。

②処理アルゴリズムの設計

①で得られた知識を基に、従来の文書処理技術を参考に処理アルゴリズムの設計を行う。このとき、一部でも類似するものは調査対象となるという特許公報調査の特性を考慮する。現時点で検討中のアルゴリズムとしては、特許公報に対し、専門家の評価結果や国際特許分類を目的変数、フレーズを説明変数とし、それらの関係を抽出することをベースとする。この時、類似度のみではなく、相違度に着目する。これらの値を計算する際に、名詞のみでなく、それに続く助詞とその役割にも重みを与える。さらに、従来手法では活用事例が少ない否定の考えを導入し、完全に偽の特許のみを排除する形での処理アルゴリズムの実現を図る。

③アルゴリズムの実装

②で検討したアルゴリズムを実装する。実装は、本研究予算において購入予定の PC 上にて実装を行う。開発言語としては、基本的に Java を用いる。

④特許公報の収集

特許公報は PDF で提供されているが、書誌情報の部分は画像データとなっている。そのため予算にて購入する PDF 処理用ソフトウェアを用い、テキストデータへの変換を行う。このデータに対し、処理を加えることになる。

３．２　2 年目以降の内容

研究 2 年目、平成 25 年度以降においては、主に以下のことを行う。

①評価実験の準備

評価実験においては、従来技術を用いたシステムとの比較評価を行う。そのために、平成 24 年度までに収集した情報を基に、従来技術を用いたシステムを構築する。

②評価実験

平成 24 年度に収集した特許候補を基に評価実験を行う。評価実験においては、同一課題に対し、提案システムと①で作成する従来技術ベースのシステムとで処理を行う。それらを比較することにより、本システムの有用性を評価する。課題としては企業と構築する独自課題のほか、NTCIR の特許検索タスクのテストコレクションを利用する。評価においては、意見収集を行った専門家や特許調査の初心者である学生に利用してもらうことにより、精度のみではなく、インタフェースやアクセシビリティといった面の評価も行う。

③評価実験結果の分析

②で得た結果を分析し、提案手法の改善点を洗い出す。基本的には、精度不足の点に対し、その理由を割り出す。また、インタフェースなどについても意見を分析、修正の方向性を検討する。

④システムの改善

③で得た結果を基に、システムの改善を進める。改善においては、詳細だけでなく、システム全体にも視点を置き、全体としての性能が悪化しないように心掛ける。この改善を進める過程においても、評価実験に参加してもらう専門家に参加してもらい、意見交換を行いつつ進める。

⑤システムの再評価

④の結果を基に、②と同様の実験を行う。その結果に対し、１回目と同様に③評価実験結果の分析、④システムの改善を行う。

４．研究成果

４．１　研究初年度の成果

１年目では、以下のことを行った。

特許の内容把握において、その特許の内容を特徴づける重要文の抽出は有意義である。重要文を提示するだけでも、特許業務にかかる人の負担を軽減することが期待できる。本研究では、(1)人手による重要部分抽出、(2)重要部分を含む文を抜き出し、N-gram の抽出、(3)抽出した N-gram を用いたフィルタ作成、(4)重要文抽出機能の評価の順番で処理を行った。評価は実務者が行い、抽出した 1025 文中 766 文（約 75%）が有用と評価された。また、こちらが提案したものと異なるフィルタの提案を受けた。この結果から、本機能は有用である可能性が示された。フィルタにより上記機能の有用性が変わること、および人手で行うのは負荷が高いと考えられることから、今後は、フィルタ構築支援機能の開発を行う必要がある。クラメールの連関係数に基づく文書類似度計算方法を提案した。これに対し、従来手法との比較検証を行った。主な内容としては、利用する語句の切り出し方法として、形態素解析の他、N-gram、辞書にある語句との最長一致、文字種区切りの４つの方法を用い、文書間の類似度計算には、クラメールの連関係数のほか、文書ベクトルによる方法も用いた。また、重要文として抽出した文書のみを用いる場合、重要文以外を用いる場合、両方を用いる場合についても考慮した。これらの組み合わせにより文書類似度を、同一の文書集合に対して算出し、実務者側で行った人の評価と比較した。人による評価は、文書を４段階でランク付けし、各ランクの類似度の平均の変動を調べた。傾向として良い結果となったのは、全文から得た形態素と文書ベクトルとの組合せ、全文から得た 2-gram と文書ベクトルとの組み合わせであった。クラメールに対しては、類似度の照合結果が偏りやすいことが示された。そのため、クラメールの連関係数を用いるのは、相違度で利用するなど再度の検証が必要である。［学会発表①②］

４．２　研究２年目の成果

２年目においては、次のことを行った。

平成 25 年度の一つ目の成果として、文書類似度計算手法の検討があげられる。本研究では、特許中の語句とブロックタグのペアの出現数を要素とする文書ベクトルを作成し、それらの比較を行う手法を検討した。語句としては、形態素、N-gram、辞書中の語句、文字種区切り、ベクトル間の類似度計算としては、Cos 類似度、クラメールの連関係数を用いるパターンを用意し、その有用性を検討した。結果として、N-gram と Cos 類似度の組み合せの結果が最良であった。

二つ目の成果として、文書分類の基盤技術となりうる決定木手法の向上を検討した。本研究では、決定木手法における子ノード作成時にクラスタリングを併用することで、決定木による分類精度の向上を図ることを試みた。手法としては、リーフノードに対し、クラスタ分析による分割が可能な場合は、さらに分割することにより、複数の属性値の関係を反映した分割を行う事を実現した。プログラムを用いて作成したデータに基づく評価の結果、従来手法との有意な差は得られなかったが、有用に働く可能性を示すことはできた。［学会発表④］

三つ目の成果として、専門家が特許に付与する課題分類、手段分類の推定支援を行う枠組みを検討した。本研究では、すでに課題分類、手段分類が与えられた特許における、語句とブロックタグのペアの出現傾向と、新たな特許の語句とブロックタグのペアの出現傾向を比較することで、課題分類、手段分類の推定を行う。文書間の比較については、一つ目の成果も活用した。研究協力者と連携した評価において、上位 10 位以内に正答が含まれる確率が６割を超えた。この結果から本手法の有用性を示すことができた。［学会発表⑤⑥］

４．３　研究最終年度の成果

３年目では、以下のことを行った。

権利調査などにおける特許公報処理支援を行うために、特許が解決しようとする課題とその手段の候補を推定する手法を提案した。本手法では、大分類と小分類の組み合わせから表現された課題分類と手段分類を、専門家が事前に行った課題分類・手段分類の抽出結果をもとに推定する。専門家の協力のもとに行った評価実験においては、課題分類・手段分類の組については、10 位以内に正答が含まれる割合は最大で、大分類小分類両方の場合は約 17%、大分類のみの場合は約 49%であった。今後の課題としては、より大規模なデータでの実証実験、語句の切り出し方やブロックタグの利用などによる各種情報抽出方法の検証と改善、計算量の削減などがあげられる。[学会⑦⑨]

また、2 年目で検討した決定木手法の改善を進めた。2 年目の手法は訓練例に着目したが、3 年目では、K-NN をすべての子ノード生成時に用いることで、より決定木の精度向上を図った。具体的には与えられた集合を K-NN で複数に分割し、その分割した集合ごとに子ノードを構築することを試みる。これにより、集合の分割基準が明確になり、訓練例の分割が進む事が期待できる。オープンデータを用いた評価の結果、従来手法との有意な差は得られなかったが、有用に働く可能性を示すことはできた。[学会発表⑧]

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）
該当無し

〔学会発表〕（計　9 件）
① 樽松理樹、クラメールの連関係数を利用した特許公報検索システムの構築、平成 24 年度電気関係学会東北支部連合大会、2012 年 8 月
② 樽松理樹、クラメールの連関係数を援用した類似文書検索システムの提案、第 11 回情報科学技術フォーラム、2012 年 9 月
③ 樽松理樹、クラメールの連関係数を援用した類似文書検索の評価、第 12 回情報科学技術フォーラム、2013 年 9 月
④ 樽松理樹、藤田ハミド、A Framework for Integrating a Decision Tree Learning Algorithm and Cluster Analysis、12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques、2013 年 9 月
⑤ 樽松理樹、専門家による抽出結果を用いた特許公報からの課題手段推定支援手法の提案、人工知能学会 第 69 回 言語・音声理解と対話処理研究会(SIG-SLUD)、2013 年 12 月
⑥ 樽松理樹、専門家による抽出結果を用いた特許公報からの課題手段推定支援手法の提案、2014 年度人工知能学会（第 28 回)、2014 年 5 月
⑦ 樽松理樹、課題と手段の類似度に基づく特許分類支援システムの提案、第 13 回情報科学技術フォーラム、2014 年 9 月
⑧ 樽松理樹、羽倉淳、藤田ハミド、A Framework for Improvement a Decision Tree Learning Algorithm Using K-NN、IEEE 13th International Conference on Intelligent Software Methodologies, Tools and Techniques、2014 年 9 月
⑨ 樽松理樹、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 年 5 月

〔図書〕（計　0 件）
該当無し

〔産業財産権〕
○出願状況（計　0 件）
該当無し

○取得状況（計　0 件）
該当無し

〔その他〕
ホームページ等
該当無し

６．研究組織
(1)研究代表者
樽松理樹（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