E-018

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

A Framework of a Patents Classification System Based on Term Frequency in Patent Journal Blocks

樗松理樹†
Masaki Kurematsu

１．はじめに

特許公報[1]は，代表的な知的財産情報であり，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」「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[2][3][4]が提案されている．その多くは，特許情報プラットフォーム [5]に代表されるような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，表層情報レベルで処理している．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課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が必要である．

一方，実務作業に目をむければ，特許公報は膨大であり，すべてを調べることは難しい．本研究の研究協力者であり，企業内の知的財産部門で特許公報を取り扱っている専門家は，その特許が述べている課題と手段を分類し，比較対象となる特許と課題および手段が類似しているものからチェックしている．これにより，特許公報の内容把握にかかる時間の軽減を図っている．しかし，特許公報が膨大であることから，特許が対象とする課題と手段の分類も大量の負荷や労力が必要となっている．

以上の背景から，著者は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[6][7]に取り組んできている．これまでの手法は，専門家が課題・手段を分類した特許中の出現語句の類似度から課題や手段の推定を試みてきた．ここで専門家とは，企業などにおいて特許処理に携わっている実務者を意味する．しかし，一つの特許の影響が強く，精度は不十分である．そのため，本稿では新たに，専門家が課題・手段を分類した特許の要約から分類推定に有用な語句情報を取り出し，それを利用することで課題，手段の分類の推定を試みる手法を提案する．

２．特許処理

2.1 特許公報の構造

本研究で対象とする特許公報は，フロントページと明細書から構成される[1]．フロントページには，発明の名称，出願人，発明者，要約，国際特許分類（IPC），FI（File Index），F タームなどが記載されている．IPC は発明の技術内容に応じた世界共通の特許分類の記号であり，一つの特許には複数ついていることが多い．FI は IPC をさらに分類したものであり，日本の独自の分類である． F タームは審査官が審査に利用する分類記号であり，FIを技術的範囲に分け、複数の件点から分類したもの[4][5]である．明細書には，特許請求の範囲，発明の属する技術分野，発明が解決しようとする課題，課題を解決するための手段などが記載されている．フロントページおよび明細書に記載されている内容については，【】で囲まれた**ブロックタグ**により，それが何について述べている部分かが明確になっている．代表的なブロックタグとしては，【特許請求の範囲】【技術分野】【背景技術】【先行技術文献】【特許文献】【発明の概要】【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】【発明を実施するための形態】などがある．

IPC や FI，F タームは，特許の分類を端的に示していることから，課題や手段の推定に利用できると考えられる．しかし，これらの分類は，請求項の内容によって付与されている点，これらの分類と実務者の考える分類と相違がある点，分類の付与が人によって異なる点，改訂によってコードが変わる点などから，IPC や FI，F タームのみでの課題や手段の把握は困難である．そのため，専門家は独自の分類を付与している．しかし，専門家間でも意見が異なる場合があり，これらの付与支援は大きな課題である．

2.2 特許の課題と手段

専門家は，特許に対し，独自の分類を付与している．代表的なものとして，その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対するものがあげられる．それぞれに対し，端的に概要を示す語句を与えている．以後，課題の分類を示す語句を**課題分類**，手段の分類を示す語句を**手段分類**と呼ぶ．課題分類と手段分類は，それぞれ大分類・小分類の組み合わせで示される．今回協力いただいた専門家は，課題分類に対し，大分類を 13 種類，小分類を 62 種類設定し，手段分類については，大分類 13 種類，小分類 33 種類を設定している．実際に利用する際には，課題・手段とも大分類から 1 つ，小分類から 1 つ選択している．たとえば，課題の大分類として「作動性能」を，小分類として「安定性」を定義し，手段の大分類として「制御装置」，小分類として「共通」を与える．

これらの分類を用いることで，自分たちの視点に基づき，権利調査の対象としての各特許の重要性を判断し，重要性の高いものから特許の確認を行うことができる．

本研究では，新規に与えられた特許に対し，過去に処理された特許をもとに，これらの課題分類・手段分類を抽出することが目的である．

†岩手県立大学ソフトウェア情報学部

3. 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

3.1 手法概要

本提案システムの概要を図1に示す．本システムは大きく「分類出現語句情報抽出部」「文書ベクトル変換部」「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分野，手段分野が付与された特許の要約から，分類ごとにその分類を特定するのに有用と思われる分類出現語句情報を抽出する．「文書ベクトル変換部」では，分類出現語句情報をもとに，分類済み特許，対象となる新規特許を文書ベクトルに変換する．「分類推定部」では，文書ベクトルや文書ベクトル間の類似度をもとに課題分類，手段分類の候補を推定する．以降で各部分の説明を加える．

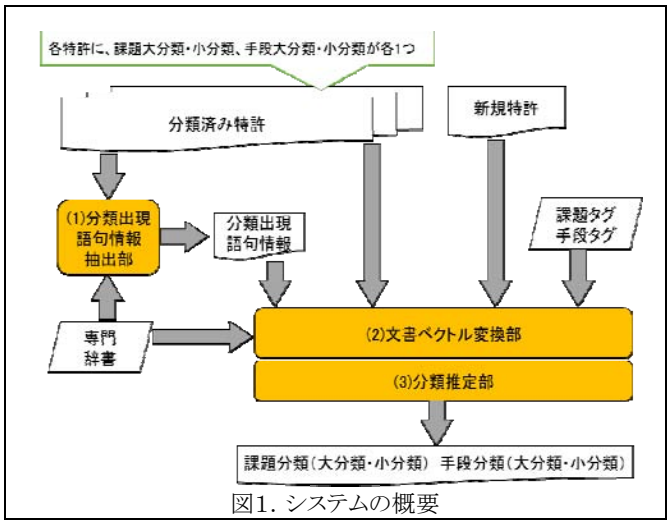

図1．システムの概要

3.2 分類出現語句情報抽出部

専門家が分類した特許公報の要約から，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から，次の方法で課題に関する文と手段に関する文を抽出する．

要約は基本的に以下のような構成となっている．

＜文章 1＞【…(課題｜目的)】＜文章 2＞

【…(手段｜構成)】＜文章 3＞

ここで、【】はブロックタグであり，前者は課題か目的で終わる，後者は手段か構成で終わることを意味している．たとえば，前者には【発明が解決しようとする課題】が合致し，後者には【課題を解決するための手段】が合致する．上記の文章 2 に含まれる文を，課題に関する文，文章 3 に含まれる文を，手段に関する文とする．

(2) 語句候補の抽出

上記のブロックに含まれる文から，(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を抽出する．形態素列としては，　"名詞-サ変"，"名詞-サ変"＋"名詞-接尾"，"名詞-サ変"＋"名詞-サ変"，"名詞-サ変"＋"名詞-一般"，"名詞-一般"，"名詞-一般"＋"名詞-接尾"，"名詞-一般"＋"名詞-サ変"，"名詞-一般"＋"名詞-一般""名詞-形容動詞語幹"，"名詞-形容動詞語幹"＋"名詞-サ変"，"名詞-形容動詞語幹"＋"名詞-一般"である．ここで＋は連続していることを意味する．

英字列，カタカナ列は 2 つ以上の英字またはカタカナの並びである．

またここで用いる専門辞書とは，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的な言葉（代表語）が与えられている．特許中では語句が出現することが多いが，語句を代表語に置き換える事で，表記の揺らぎの吸収を図っている．

これらの切り出し方は，これまでの研究成果[6][7]に基づき決定した．

(3) 重さの決定

以上で抽出した語句 t_i について，分類 C_j における出現数 $tf(t_i,C_j)$，および出現文数 $sf(t_i,C_j)$を求める．さらに，式(1)および式(2)を用いて，語句 t_i の分野 C_j における出現比率 $wtf(t_i,C_j)$，および語句 t_i の分野 C_j における出現文比率 $wsf(t_i,C_j)$求める．これらの値が高いほど，語句 t_i は分野 C_j を特定するのに有用であると考える．

$$wtf(t_i,C_j)=\frac{tf(t_i,C_j)}{tf(t_i,C_*)} \qquad 式(1)$$

$$wsf(t_i,C_j)=\frac{sf(t_i,C_j)}{sf(t_i,C_*)} \qquad 式(2)$$

(4) 語句の絞込み

上記で求めた語句のうち，出現傾向が特定の分類に偏っているものが有用であると考え，求めた重みが，2/分類数以上のもののみに絞り込む。ここで，2/分類数としたのは，出現傾向が一様分布になっている場合，1/分類数と考えられるため，単純にその 2 倍以上と設定している．また，出現回数が 2 回以上もの，および重みが 0.3 以上のものに絞込みを行っている．これらの値はパラメータ設定であり，変更可能である．

(5) 語句組の生成

さらに上記で切り出した語句について，同一ブロックに出現する異なる語句について，語句組を作る，すなわち共起関係を求める．今回，語句組の重さは，それぞれの語句の重さの積としている．以後，語句と語句組をまとめた，語句(組)と表記する

以上で求めた，(分類 C_j，語句(組)t_i，出現数 $tf(t_i,C_j)$，出現文数 $sf(t_i,C_j)$，出現比率 $wtf(t_i,C_j)$，出現文火散る $wsf(t_i,C_j)$)の集合を，分類出現語句情報とする．

3.3 文書ベクトル変換部

得られた分類出現語句情報を用いて，各特許を以下の方法で文書ベクトルに変換する．なお，課題と手段は別々に処理するため，課題推定用，手段推定用それぞれの文書ベクトルを構築する．

(1) 対象とする範囲の抽出

特許の構造に着目し，課題推定用には課題に関係するブロック，手段推定用には手段に関係するブロックを抽出する．これは，専門家が権利調査する際に特許のすべてに着目していないという知見に基づいている．このブロックをわけるために，これらはブロックタグに対する照合パターンを“.*課題】”というような正規表現で示した課題タグ，手

段タグを用いる．ブロックに与えられたブロックタグと，課題タグ，手段タグをそれぞれ照合し，条件を満たしたブロックタグをもつブロックを抽出する．

(2) 文書ベクトルへの変換

抽出したブロック中に出現する文章から，分類出現語句情報抽出の(2)と同じ方法で語句を取り出し，それらから語句組を作成する．

それら語句(組)と分類出現語句情報とを照合し，分類 C_j の重さ $V(C_j)$を求め，それらを要素とする文書ベクトルを構築する．分類毎の重さは，ナイーブ・ベイズ[8]の考えを基に，表 1 に示す方法で求める．ここで，$tf(t_i,d_k)$は対象特許 d_k における語句(組) t_i の出現回数，$sf(t_i,d_k)$は対象特許 d_k における語句(組) t_i の出現文数を示す．また，tp_i は対象特許 d_k 中に現れる語句(組)，tn_i は対象特許 d_k 中に現れない語句(組)を示す．

表1：重みの計算手法

手法①	$V(Cj)=\Sigma\,(tf(t_i,C_j)\times tf(t_i,d_k))$
手法②	$V(Cj)=\Sigma\,(tf(t_i,Cj))$
手法③	$V(Cj)=\Sigma\,(\log(10+wtf(t_i,C_j))\times tf(t_i,d_k))$
手法④	$V(Cj)=\Sigma\,(\log(10+wtf(tp_i,C_j)))+\Sigma\,(\log(10+1-wtf(tn_i,Cj))$
手法⑤	$V(Cj)=\Sigma\,(sf(t_i,C_j)\times sf(t_i,d_k))$
手法⑥	$V(Cj)=\Sigma\,(sf(t_i,Cj))$
手法⑦	$V(Cj)=\Sigma\,(\log(10+wsf(t_i,C_j))\times tf(t_i,d_k))$
手法⑧	$V(Cj)=\Sigma\,(\log(10+wsf(tp_i,C_j)))+\Sigma\,(\log(10+1-wsf(tn_i,Cj))$

3.4 分類推定部

分類推定部では，3.3 で作成した文書ベクトルをもとに分類を推定する．

一つ目の方法では，文書ベクトルの各値を比較し，最大値を持つ要素の分野を推定結果とする．

二つ目の手法で，文書ベクトル間の類似度を元に分類を推定する．この手法では，はじめに，対象特許の課題，手段の文書ベクトル V と，分類済み特許の課題，手段の各文書ベクトル Wi 間の Cos 類似度[9]を，式(3)を用いて算出する．ここで，v_j，$w_{i,i}$ は，それぞれ V，Wi の j 番目の要素の値である．

$$sim(V,W_i)=\frac{\sum(v_j)(w_{i,j})}{\sqrt{\sum(v_j)^2}\sqrt{\sum(w_{i,j})^2}} \qquad \text{式(3)}$$

このとき，同じ課題分類，手段分類が付与された文書ベクトルが複数あることから，類似度の平均値により評価する．以上の方法で求めた値の降順で候補を提示する．

4 評価実験

4.1 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 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．図 2 にスクリーンショットを示す．形態素解析としては，lucene-gosen-4.0.0-naist-chasen [8]を用いている．

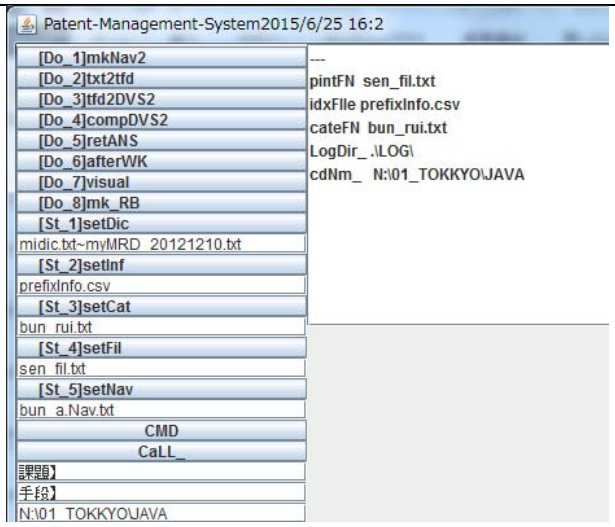

図 2 システムのスクリーンショット

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 247 件に対し，課題分類と手段分類の推定を試みる．実験の流れは以下の通りである．

(1) 特許を一つ取り出す．
(2) 残りの特許群から，(1)で取り出した特許の分類推定を行う．
(3) 推定を行っていない特許があれば，(1)に戻る．

本システムにおいては，課題および分類の検索範囲を限定するためにそれぞれタグを与える必要がある．今回は，課題タグとしては“.*課題】”，すなわち，“課題】”で終わるブロックタグ，手段タグとしては“.*手段】”，すなわち，“手段】”で終わるブロックタグを与えた．

また今回の実験においては，文書ベクトルの変換に利用する語句は，その特許での出現回数が 2 回以上のもののみとした．これは，1 回のみ出現している語句の重要度は低いという仮説と，計算量を考慮しての判断である．

評価としては，専門家が付けた分類を正解とし，それが何番目に抽出されたかにより評価する．評価については式(4)によって求めた評価得点 $E(x)$を用いる．ここで，x は処理した特許を示す．$Rank(x)$は，正答の順位，$Even(x)$は，同点の分野数，All_Cat は全分野数である．端的に言えば，全分野候補における，正解と同等以上の評価を受けた分野の割合となる．また，正答が見つからない場合を明確にするため，そのような場合には，$E(x)$の値は 2 とする．なお，文書ベクトルにおいて，正答の分類の値が 0 の場合も正答を見つけなかったものと判断している．

$$E(x)=\frac{Rank(x)+Even(x)}{All\_Cat} \qquad \text{式(4)}$$

評価する分類の組み合わせについて，「課題大分類・課題小分類」「課題大分類」「手段大分類・手段小分類」「手段大分類」について評価する．

4.2 分類出現語句情報抽出結果

分類出現語句情報抽出によって，手法①②で利用する $tf(t_i, C_j)$については，2,676 個の語句と 199,998 個の語句組が得られた．また，手法③④で利用する $wtf(t_i, C_j)$については，1,545 個の語句と 195,526 個の語句組が得られた．手法⑤⑥で利用する $sf(t_i, C_j)$については，1,199 個の語句が，手法③④で利用する $wsf(t_i, C_j)$については，2,204 個の語句と 199,998 個の語

句組が得られた．これらの概要を表 2a，2b に示す．表 2a,b において，“型”は 3.3 で述べた各値を示す．“種類”は，語句または組を，“区分”は，課題または手段を示す．“分類”の“大小”は，大分類と小分類両方を，“小”は小分類を示す．これらの組み合わせで，何に対するものかを示している．たとえば，表 2a の 2 行目は，課題・大分類小分類に対する出現頻度を元に分類出現語句情報として抽出された語句の情報であることを意味している．“分野数”は，該当するものに含まれる分野のパターン数である．たとえば，表 2a の 2 行目であれば，分類出現語句情報として抽出された語句に関連付けられた課題・大分類小分類のパターン数は 32 であることを示す．個数の“平均”と“SD”は分野パターン毎の語句や組の数の平均と標準偏差を示す．

分類出現語句情報抽出に利用した 249 件の特許に付与された課題・大分類小分類は 53 パターン，課題・大分類は 12 パターン，手段・大分類小分類は 33 パターン，手段・大分類は 8 パターンであるため，いくつかの分野については分類出現語句情報が落ちてしまっていることがわかる．

表 2a：分類出現語句情報（出現回数）

型	種類	区分	分類	分野数	個数 平均	個数 SD
tf	語句	課題	大	12	26.1	19.0
tf	語句	課題	大小	32	2.4	1.7
tf	組	課題	大	10	272.0	336.7
tf	組	課題	大小	5	3.6	2.0
tf	語句	手段	大	8	196.5	159.8
tf	語句	手段	大小	32	22.3	28.6
tf	組	手段	大	8	23388.8	30264.5
tf	組	手段	大小	25	406.0	826.8
wtf	語句	課題	大	11	4.8	3.5
wtf	語句	課題	大小	22	1.8	1.2
wtf	組	課題	大	9	236.2	276.0
wtf	組	課題	大小	5	3.6	2.0
wtf	語句	手段	大	8	100.3	88.9
wtf	語句	手段	大小	32	20.3	26.1
wtf	組	手段	大	8	22907.8	29878.6
wtf	組	手段	大小	25	404.8	826.1

表 2b：分類出現語句情報（出現分数）

型	種類	区分	分類	分野数	個数 平均	個数 SD
sf	語句	課題	大	12	17.8	16.0
sf	語句	課題	大小	17	2.4	1.2
sf	語句	手段	大	7	114.3	102.8
sf	語句	手段	大小	14	10.4	10.9
wsf	語句	課題	大	10	272.0	336.7
wsf	語句	課題	大小	22	1.8	1.2
wsf	組	課題	大	12	16.0	10.1
wsf	組	課題	大小	31	2.4	1.8
wsf	語句	手段	大	8	155.4	140.8
wsf	語句	手段	大小	32	22.1	28.5

表 2b：分類出現語句情報（出現分数）

型	種類	区分	分類	分野数	個数 平均	個数 SD
wsf	組	手段	大	8	23388.8	30264.5
wsf	組	手段	大小	25	406.0	826.8

4.3 課題分類・手段分類推定結果

表 3a～3d に実験結果の概要を示す．各表において，“手法” は，表 1 に示した手法である．Tt と Ts は比較のために行った従来手法である．この手法では，特許を，出現する語句とその出現頻度（Tt の場合は語句，Ts の場合は文数）を要素にする文書ベクトルに変換し，それらの間の Cos 類似度を求める．対象となる特許ともっとも類似度の高い特許の分類を回答とするものである．“推定”において，単は，3.4 で述べた 1 つ目の手法，すなわち，文書ベクトルの重さで推定することを意味し，複は，述べた 2 つ目の手法，複数の文書ベクトルとの比較に基づく手法で推定することを意味する．“平均”と“SD”はそれぞれ評価得点の平均値と標準偏差（SD）を示している．平均が小さいほど高い順位に正答が来ることを意味し，SD が小さいほど，値が固まっていることを意味する．平均については最小値を太字にしている．また，“≦0.2”，“≦0.5”，“≦1.0”はそれぞれ評価得点が，0.2 以下，0.5 以下，1.0 以下の個数を示している．正答は評価得点を問わずに正答を得た数であり，Miss は正答を見つけられなかった数を示している．なお課題・大分類小分類の最小値は，0.03，課題・大分類の最小値は 0.08，手段・大分類小分類の最小値は，0.03，手段・大分類の最小値は 0.13 となる．

表 3a：実験結果（課題大分類）

手法	推定	平均	SD	≦0.2	≦0.5	≦1.0	正答	Miss
①	単	0.86	0.30	14	44	143	249	0
②	単	0.87	0.27	5	40	153	249	0
③	単	**0.6**	0.27	25	95	248	248	1
④	単	0.86	0.27	25	36	249	249	0
①	複	1.03	0.19	6	11	28	249	0
②	複	0.92	0.30	3	59	67	249	0
③	複	1.01	0.23	9	16	46	248	1
④	複	1.04	0.16	6	7	51	249	0
Tt		1.04	0.17	4	10	24	249	0
⑤	単	0.89	0.30	17	38	120	249	0
⑥	単	0.94	0.26	8	26	97	249	0
⑦	単	0.77	0.27	10	49	201	249	0
⑧	単	0.77	0.41	58	71	117	249	0
⑤	複	0.98	0.28	22	24	45	249	0
⑥	複	1.03	0.18	6	9	37	249	0
⑦	複	1.02	0.20	7	14	39	249	0
⑧	複	1.02	0.21	9	13	38	249	0
Ts		1.02	0.20	6	14	38	249	0

表 3b:実験結果(課題大分類・小分類)

手法	推定	平均	SD	≦0.2	≦0.5	≦1.0	正答	Miss
①	単	0.28	0.12	60	249	249	249	0
②	単	0.12	0.06	230	249	249	249	0
③	単	**0.11**	0.14	223	248	248	248	1
④	単	0.41	0.09	19	249	249	249	0
①	複	0.97	0.21	11	13	25	249	0
②	複	0.98	0.18	8	9	21	249	0
③	複	1	0.15	4	5	20	248	1
④	複	0.97	0.21	12	13	27	249	0
Tt		0.76	0.17	11	17	249	249	0
⑤	単	0.26	0.13	36	248	248	248	1
⑥	単	0.19	0.15	110	248	248	248	1
⑦	単	0.08	0.03	242	249	249	249	0
⑧	単	0.57	0.14	18	26	249	249	0
⑤	複	0.97	0.20	8	14	21	249	0
⑥	複	0.98	0.18	8	11	18	249	0
⑦	複	0.97	0.20	8	13	30	249	0
⑧	複	0.98	0.19	10	10	23	249	0
Ts		0.79	0.11	5	7	249	249	0

表 3c:実験結果(手段大分類)

手法	推定	平均	SD	≦0.2	≦0.5	≦1.0	正答	Miss
①	単	0.96	0.23	0	32	143	249	0
②	単	0.73	0.36	32	87	199	249	0
③	単	0.78	0.31	20	58	249	249	0
④	単	1.1	0.10	0	2	27	249	0
①	複	0.92	0.36	39	43	82	249	0
②	複	0.85	0.43	58	66	93	249	0
③	複	1.1	0.11	1	3	22	249	0
④	複	0.88	0.39	25	70	81	249	0
Tt		0.77	0.32	16	74	249	249	0
⑤	単	**0.66**	0.36	69	83	249	249	0
⑥	単	0.94	0.14	0	13	249	249	0
⑦	単	0.71	0.40	63	86	205	249	0
⑧	単	1.1	0.10	0	1	20	249	0
⑤	複	1.02	0.21	1	19	70	249	0
⑥	複	1.09	0.12	0	5	38	249	0
⑦	複	1.09	0.12	2	3	33	249	0
⑧	複	1.03	0.18	1	14	81	249	0
Ts		0.82	0.30	21	50	249	249	0

表 3d:実験結果(手段大分類・小分類)

手法	推定	平均	SD	≦0.2	≦0.5	≦1.0	正答	Miss
①	単	0.34	0.24	91	174	249	249	0
②	単	0.34	0.20	82	184	249	249	0
③	単	0.53	0.25	26	91	248	248	1
④	単	0.89	0.26	8	30	249	249	0
①	複	0.94	0.24	7	28	39	249	0
②	複	0.97	0.22	12	15	32	249	0
③	複	0.99	0.17	5	10	18	249	0
④	複	0.94	0.25	12	27	38	249	0
Tt		0.75	0.23	21	28	248	248	1
⑤	単	0.27	0.17	89	248	248	248	1
⑥	単	0.30	0.16	64	248	248	248	1
⑦	単	**0.11**	0.19	211	234	249	249	0
⑧	単	0.83	0.33	32	43	249	249	0
⑤	複	1	0.14	4	6	27	249	0
⑥	複	1	0.15	5	7	29	249	0
⑦	複	1	0.17	7	8	24	249	0
⑧	複	1	0.15	6	6	26	249	0
Ts		0.81	0.18	10	12	248	248	1

4.4 課題分類・手段分類推定の評価・考察

手法①から手法④と Tt が語句の出現数に基づいており，手法⑤から手法⑧と Ts が語句の出現文数に基づいている．これらを比較した場合，表 3 で示すように，従来手法の Tt, Ts に対し，提案手法のほうが良い結果を得た場合がある．よって，本提案手法は従来手法よりも改善できたと考えられる．

一方，分類出現語句情報が落ちているのにもかかわらず，正答を見つけている場合がある．これは，ベクトルのみの場合は，要素の値が 0 の場合も正答とみなしていることが原因である．この点については方針の見直しが必要である．また Miss となっているものは，ほかに同じ分野に割り当てられている特許がなかったことが起因している．このような，いわば欠損値への対策も必要となると考えられる．

手法①から手法④を比較した場合，課題の場合は，手法④が，手段の場合は手法①②がよい結果となっている．これは，分類出現語句情報の偏りが影響していると考えられる．数や種類が多い場合，手法③や手法④では分類の重さが近くなり，結果，識別率が落ちるものと考えられる．特に手法④については，今回のように当てはまらない場合が多いと，当てはまらなかった場合の値が大きくなり，より分類の重さが均衡し，識別率が落ちていると考えられる．

手法⑤から手法⑧についても手法①から手法④と同等の傾向が見られる．また，手法①から手法④と比較した場合，結果としては似た傾向が出ている．基本的に手法①と手法⑤，手法②と手法⑥，手法③と手法⑦，手法④と手法⑧とでは利用する値が，出現回数か出現文数かの違いであり，計算方法は同じである．類似した結果が出ていることから，出現回数と出現文数が似た傾向にあると思われる．これは，今回分類語句情報を要約から抽出したためと思われる．本文などを利用した場合は，傾向が異なる可能性はある．

推定手法においては，文書ベクトル単体のほうが，複数の文書ベクトルを利用する方法よりも良い結果が出ている．これは複数文書ベクトルを比較することで，分野ごとの差が吸収されてしまうこと，また一つの文書ベクトルから影響が大きい場合があるためと考えられる．一方で，複数の文書ベクトルを利用すれば，分類語句情報に含まれない分類の発見が期待できる．

分類としては，大分類と小分類の両方を推定した場合のほうが，大分類のみを推定した場合より，良い結果を得た．これは，表 2a,b で示したように，大分類のみのほうが分類に利用する語句や語句組のパターンが少なく，それぞれの数が多いことから，分類の差が生じにくくなったためと予想される．

今回従来手法の改善については行えたが，評価得点が 0.2 以下，すなわち全候補数の上位 20%に正答を得た場合は少ない．また，対象によって手法の差が出ており，全てに有用な方法が決定できていない．今後は今回の実験結果をさらに解析し，これらの問題点を改善する必要がある．本提案手法は，分類出現語句情報に依存するところが大きい．そのため，今後は，抽出方法の改善に重点を置く．具体的には，今回の結果より，出現文数に基づく方法もある程度有用であると評価できたことから，出現語句情報の大きさや分野の網羅する範囲などを考慮し，この方法を優先する．現在，出現数の制約をつけているため，これを変更し，より有用な情報の抽出を試みる．また，組については，現在，語句間で共通部分がある場合も組としてみている．これが組の数を大きくしている主要因である．そのため，組の作り方を見直すとともに，文単位ではなく，ブロック単位での出現に着目する．また，現在のような語句とその組単位ではなく，複数の語句の出現によるルール作成も視野にいれる．この方法は全ての語句を対象とすると膨大になると考えられるため，まずは専門辞書に出現する語句を利用することを考える．また，システムの性質上，既存の結果に依存するところが大きい．この問題を解消するために，新たな結果を反映する学習機能を検討する．

5 おわりに

本稿では，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，専門家が事前に課題分類・手段分類を抽出した特許における語句や語句組の出現情報をもとに新規特許の分類を推定する．専門家の協力のもとに行った評価実験においては，課題分類・手段分類の組について従来手法よりも向上することができた．今後は，語句の切り出し方や計算方法の再検討，学習機能の追加などによる改善を進める予定である．

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

[1] 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）
[2] 寺岡岳夫：特許情報検索の現状と今後，Japio Year Book 2010，pp.166 - 169（2010）
[3] 谷川英和：特許と情報学－特許実務における情報学の貢献と研究者等の特許活動－，情報処理学会，Vol.54，No.3，pp.192 - 199（2013）
[4] 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）
[5] 工業所有権情報・研修館：特許情報プラットフォーム，https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage（2015/6/22 アクセス）
[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，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)，pp.49-54，（2013）
[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，pp.237-242，（2014）
[8] Domingos, Pedro and Michael Pazzani： "On the optimality of the simple Bayesian classifier under zero-one loss". Machine Learning, Vol.29，pp.103–137 (1997)
[9] 北 研二，津田和彦，獅々堀正幹：“情報検索アルゴリズム”， 共立出版 (2002)
[10] Lucene-gosen，https://code.google.com/p/lucene-gosen/（2015/6/24 アクセス）