E-016

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

A Framework of a Patents Classification System Based on the Similarity of Tasks and Methods

樗松理樹†
Masaki Kurematsu

1. まえがき

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]が提案されてきた．その多くは，特許電子図書館(IPDL)サービス[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を行っている．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を確立することが依然として求められている．

一方，実務作業に目をむければ，すべての特許公報を読むことは難しい．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，特許で取り組む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，本論文では，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を提案する．本提案手法は，専門家が課題・手段を分類した特許と，対象となる特許との類似度を，特許構造を考慮して求める．この値をもとに，課題，手段の分類の推定を試みる．なお，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．

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（テーマ）に分け、複数の件点から分類したものである [4][5]．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．

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

先に述べたように，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．今回協力を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，その組み合わせパターンを 66 種類利用している．一つの大分類における小分類の数は平均 5.5 種類である．一方，手段分類については，大分類 13 種類，小分類 33 種類，組み合わせパターンを 40 種類を利用している．一つの大分類における小分類の数は平均 5 種類である．その一部を表 1 に示す．

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．

表1：課題分類と手段分類の一部

区分	大分類	小分類
課題	作動性能	安定性
	作動性能	信頼性
	構造性能	耐久性向上
手段	制御装置	共通
	その他の手段	その他

†岩手県立大学ソフトウェア情報学部

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

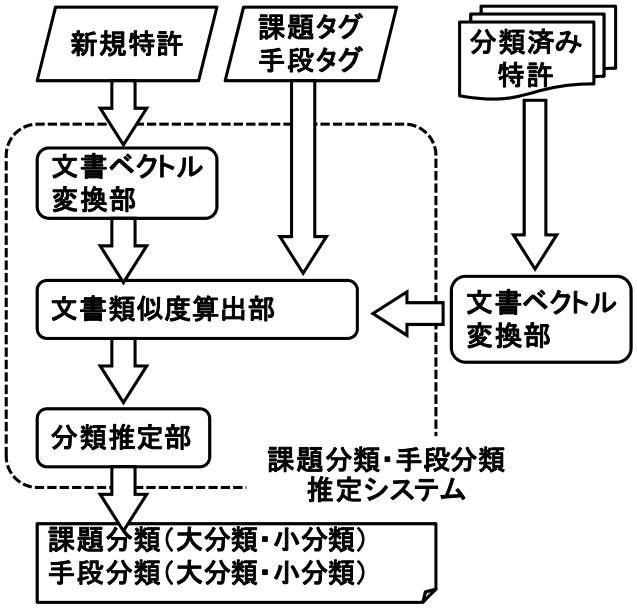
3. 特許構成を考慮した文書類似度に基づく特許からの課題分類・手段分類推定システム

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「文書ベクトル変換部」「文書類似度算出部」「分類推定部」からなる．処理手順は次に示す通りである．

入力としては，新規特許と課題タグ，手段タグを与える．課題タグ，手段タグとは，課題や手段を抽出する際に注目するブロックタグを限定するために用いる．詳細は後述する．

システムは最初に「文書ベクトル変換部」において、新規特許を文書ベクトルに変換する．次に同じ方法で文書ベクトル化された課題・手段分類済み特許との類似度を「文書類似度算出部」で求める．ここで，課題・手段分類済み特許とは，専門家が課題分類および手段分類を付与した公開特許である．以後，分類済み特許と呼ぶ．なお分類については，それぞれ大分類 1 つと小分類1つからなる組が課題，手段に対し一つずつ与えられている．最後に求まった類似度をもとに「分類推定部」で課題，手段の各候補を出力する．



※分類済み特許には，課題大分類・小分類および手段大分類・小分類の組が一つずつ与えられている．

図1．システムの概要

3.2 文書ベクトル変換部

文書ベクトル変換部では，次の方法で各特許文書を文書ベクトルに変換する．

1. 各文のブロックタグが，【発明の名称】【要約】および【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】などの明細書に含まれるもの場合，以下の文字列を切り出す．
   ①形態素，②連続するカタカナ，③後述の専門辞書に登録されている語句（代表語も含む）
2. ブロックタグと切り出した文字列をペアにしたものを作成する．これをパターンと呼ぶ．
3. それぞれのパターンの出現回数を要素の値とする文書ベクトルを作成する．

上記の 1 で利用する専門辞書とは，専門家によって構築された辞書であり，語句とそれに対する代表語が記載されている．ここで代表語とは，その語句の概念を示す代表的な言葉である．

3.3 文書類似度算出部

文書類似度算出部では，次の方法で文書ベクトル(V_1, V_2)間の類似度を算出する．基本的に，Cos 類似度[6]を用いている．類似度の算出においては，特許の構造に着目し，IPC や FI などのコード部，課題に関係する部分，手段に関係する部分の 3 ブロックにわけて処理を行う．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，課題タグ，手段タグを用いる．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与えている．各パターンにはブロックタグが含まれるため，条件を満たしたブロックタグをもつパターンのみを対象に次の方法で類似度を算出する．

1. パターンが，課題タグを満たす場合は，課題の類似度 T を式(1)で求める．ここで，$V^t{}_1$, $V^t{}_2$は，V_1, V_2それぞれの課題タグにマッチする部分であり，t_iは，$V^t{}_1$, $V^t{}_2$のいずれかまたは両方に含まれるパターンである．また，Tf(t_i , $V^t{}_1$)，Tf(t_i , $V^t{}_2$)はそれぞれ，$V^t{}_1$の t_i の出現回数，$V^t{}_2$の t_i の出現回数である．

$$T = \frac{\sum Tf(t_i, V_1^t) Tf(t_i, V_2^t)}{\sqrt{\sum Tf(t_i, V_1^t)^2} \sqrt{\sum Tf(t_i, V_2^t)^2}} \qquad \text{式(1)}$$

2. ブロックタグが，手段タグを満たす場合，手段の類似度 M を課題の類似度 T と同様に式(2)で求める． ここで，$V^m{}_1$, $V^m{}_2$は，V_1, V_2それぞれの手段タグにマッチする部分であり，t_iは，$V^m{}_1$, $V^m{}_2$のいずれかまたは両方に含まれるパターンである．また，Tf(t_i , $V^m{}_1$)，Tf(t_i , $V^m{}_2$)はそれぞれ，$V^m{}_1$の t_i の出現回数，$V^m{}_2$の t_i の出現回数である．

$$M = \frac{\sum Tf(t_i, V_1^m) Tf(t_i, V_2^m)}{\sqrt{\sum Tf(t_i, V_1^m)^2} \sqrt{\sum Tf(t_i, V_2^m)^2}} \qquad \text{式(2)}$$

分類済み特許には，課題分類，手段分類がそれぞれ一つずつ与えられている．よって，この結果を課題分類，手段分類の類似度として次の処理で分類推定を行う．

3.3 分類推定方法

3.2 で述べた方法により，新規特許に対する課題分類，手段分類の類似度を得る．分類済み特許には課題分類，手段分類が一つずつ与えられているため，同じ課題分類，手段分類に対し，複数の類似度が存在する．また，それぞれの出現数は異な

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

る．これに対し，次にあげる 4 つの方法で課題分類，手段分類の類似度を統合する．

（A） 分類 i の類似度の最大値を選択．
本手法では，単純に類似度が高いほど述べている内容は類似していると判断し，それらの関係は高いという考えをもとに最大値を選択する．

（B） 分類 i の類似度の最小値を選択する．
本手法では，最大値とは逆に，最小値をとる．それぞれの類似度としては最小値以上の値をとると考え，その値が最も高いもの，最悪の場合で，もっとも高いものを優先する．

（C） 分類 i の類似度の平均値を選択．
本手法では，同じ分類の類似度の算術平均をとる．値に幅があるため，利用頻度の高い平均値を用いる．

（D） 分類 i の類似度の CF 値を選択．
MYCIN[7]で用いられている CF 値（確信度係数）の考え方を援用し，代表値を決める．分類iの類似度として，sim1，sim2 が求まった時，以下の方法で類似度を計算する．

CF 値（sim1，sim2）＝sim1＋sim2 － sim1×sim2

3以上ある場合は，これを繰返し適用していく．

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．形態素解析としては，lucene-gosen-4.0.0-naist-chasen [8]を用いている．また特許は PDF 形式で与えられているため、PDF からテキストを取り出すために、xdoc2txt[9]を用いた．実装したシステムのスクリーンショットを図 2 に示す．なおいくつかのボタンは，本研究と直接関係ない支援機能を呼び出すものである．

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 639 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) 残りの特許群をもとに，(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．

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．また，分類については，「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．

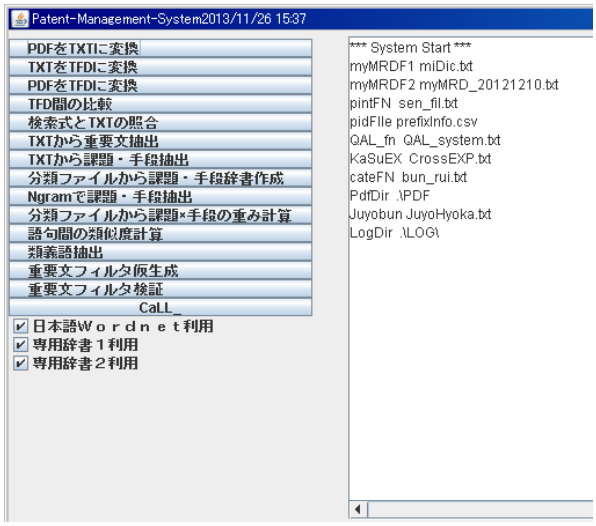


図 2 システムのスクリーンショット

4.2 実験結果

課題分類に対する結果を表 2 に示す．表 2 において，範囲の「大小」は大分類と小分類の組，「大」は大分類のみを意味する．手法は，3.3 で示した 4 つの統合方法である．また 1 位から 10 位は，正解が出現した順位を示す．なお順位は同順も許しているが，その場合，次の順位は同順分を加えた順位とする．以降は，11 位以降を示す．Errは正解となる分類を見つけることが出来なかった数である．平均は，解答に付与された値の平均である．四つの方法で利用する値が異なるため単純比較はできない．

表 2. 実験結果（課題の推定結果）

範囲	大小	大小	大小	大小	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
1 位	16	14	13	484	17	47	83	623
2 位	0	1	10	2	14	592	97	3
3 位	4	1	13	3	25	0	80	0
4 位	0	1	19	0	18	0	59	0
5 位	0	1	18	0	25	0	56	1
6 位	2	3	8	1	28	0	60	0
7 位	0	1	16	0	44	0	50	2
8 位	0	3	13	0	44	0	48	1
9 位	1	2	9	0	49	0	44	0
10 位	0	7	16	2	86	0	39	1
以降	601	590	489	132	289	0	23	8
Err	15	15	15	15	0	0	0	0
平均	0.69	0.17	0.51	0.95	0.74	0.00	0.50	0.98

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

手段分類に対する結果を表 3 に示す．各項目は，表 2 と同じである．

表 3. 実験結果（手段の推定結果）

範囲	大小	大小	大小	大小	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
1 位	14	21	21	534	24	632	67	622
2 位	2	9	16	7	10	0	119	6
3 位	1	5	8	0	23	0	70	0
4 位	2	5	19	0	57	0	108	2
5 位	1	1	11	3	61	0	125	1
6 位	4	2	21	1	97	0	106	0
7 位	2	2	14	0	133	0	35	1
8 位	4	4	16	0	234	7	9	7
9 位	5	6	12	0	0	0	0	0
10 位	3	9	14	1	0	0	0	0
以降	599	573	485	91	0	0	0	0
Err	2	2	2	2	0	0	0	0
平均	0.70	0.06	0.50	0.96	0.73	0.01	0.50	0.98

4.3 評価・考察

(1)順位についての評価・考察

実験の結果，上位 5 位に含まれる正解の割合は，課題の大分類小分類の組み合わせについては，最大値が約 3%，最小値も約 3%，平均値は約 11%，CF 値が約 77%となった．大分類のみの場合は，最大値が約 15%，最小値は 100%，平均値が約 59%，CF 値が約 98%となった．手段については，大分類小分類の組み合わせについては，最大値が約 3%，最小値は約 6%，平均値が約 11%，CF 値が約 85%であり，大分類のみの場合は，最大値が約 27%，最小値は約 98%，平均値が約 77%，CF 値が約 99%となった．また，同順の平均個数は表 5 のようになり，CF 値や最小値を用いた場合，複数のものが候補として選ばれていることが分かる． また表 4 の候補は今回利用した分類付与済み特許における分類の種類数である． 最小値や CF 値を用いる場合，正解は上位に出てくるが，同順候補を考えるとその精度は低い．特に手段大分類のみにおいては，8 種類中7種類が候補として挙げられることになり，意味がない．

表 4. 実験結果（同順の平均数）

範囲	大小	大小	大小	大小	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
課題	2.6	14.7	2.6	19.4	1.3	10.5	1.3	10.5
候補	51				12			
手段	1.9	17.3	1.9	12.9	1.2	6.8	1.2	6.8
候補	38				8			

表 4 に示すように，大分類小分類，大分類のみでは候補数が異なるため，順位を候補数で割った割合で比較した結果を，表 5，表 6 に示す．表において，範囲，手法は表 2，3 と同じである．##%の部分は，上位から##%までの 10%内に含まれる回答数を示す．例えば、20%であれば、上位 10%以上、20%以内を示す．手段の大分類は候補が 8 個のため，1 位であっても 0.125 となり，10%以下が 0 となる．傾向としては，平均値を利用した場合は全体的に分布しているのに対し，最大値は 60%以降に偏っている．また最小値については，大分類小分類の場合は，60～70%に偏っている．CF 値については，順位と同様に上位に固まっていることが分かる．

表 5. 実験結果（課題の推定結果・分布）

範囲	大小	大小	大小	大小	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
10%	37	36	96	505	17	47	83	623
20%	3	31	100	3	14	592	97	3
30%	10	42	63	3	25	0	80	0
40%	12	45	75	25	18	0	59	0
50%	26	40	62	27	25	0	56	1
60%	33	73	71	37	72	0	110	2
70%	65	372	66	23	44	0	48	1
80%	64	0	44	7	49	0	44	0
90%	117	0	47	4	86	0	39	1
100%	272	0	15	5	289	0	23	8

表 6. 実験結果（手段の推定結果・分布）

範囲	大小	大小	大小	大小	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
10%	19	37	47	543	0	0	0	0
20%	9	10	65	4	24	632	67	622
30%	16	20	64	2	10	0	119	6
40%	8	9	69	5	23	0	70	0
50%	13	28	85	13	0	0	0	0
60%	24	535	90	8	57	0	108	2
70%	35	0	90	20	61	0	125	1
80%	61	0	75	26	97	0	106	0
90%	106	0	36	3	133	0	35	1
100%	348	0	18	15	234	7	9	7

順位について，同順数を加えた順位を求め，順位の出現数をまとめた結果を表7，表 8 に示す．表 7 は，大分類小分類の結果であり，本表において，1,2 は 1 位と 2 位を意味する．また，TOP10 は上位 10 位までの数であり，その下の%は，その割合を示している．表 8 は大分類のみの結果を示しており，TOP2 は上位 2 位までの数、下の%は，その割合を示している．

結果として，課題の大分類小分類に対しては，平均値を用いた時が 21%となった．他の手法が 5%程度であることから，本手法が今回の中では最も良いと評価する．手段の大分類小分類についても同様である．課題の大分類，手段の大分類についても同様に，平均値を用いた時が 26%，27%と高くなった．

全体として，CF 値や最小値は同じような値になりやすく，個々の課題分類や手段分類を識別するのには不適切であると言える．また，最大値は，よい値を得られる場合もあるが，その差が大きい．主たる指標ではなく，二次指標としての利用が考えられる．一方，平均値は全体としてやや高い結果を得ているが，精度は低い．これは値の範囲による依存も大きいと考えられる．今回の手法の中では最も良いが，能力としては不十分であ

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

る．よって，最頻値や中央値といったこれらとは別の指標や，確率論に基づく手法，これらを統合する手法を検討する必要がある．

表 7. 実験結果（大分類・小分類の推定結果）

範囲	課題	課題	課題	課題	手段	手段	手段	手段
範囲	大小	大小	大小	大小	大小	大小	大小	大小
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
1,2	17	16	26	16	5	17	26	2
3,4	4	2	29	1	3	10	23	5
5,6	2	4	27	0	5	2	30	4
7,8	0	3	23	1	6	7	32	1
9,10	1	8	26	4	7	6	28	2
11,12	1	13	33	1	7	3	31	54
13,14	3	11	22	2	4	5	46	61
15,16	3	15	28	3	4	12	31	68
17,18	3	9	18	35	5	10	41	355
19,20	3	21	21	25	7	9	49	2
TOP10	24	33	131	22	26	42	139	14
%	4%	5%	21%	3%	4%	7%	22%	2%

表 8. 実験結果（大分類のみの推定結果）

範囲	課題	課題	課題	課題	手段	手段	手段	手段
範囲	大	大	大	大	大	大	大	大
手法	最大	最小	平均	CF 値	最大	最小	平均	CF 値
1 位	4	23	71	3	11	0	53	1
2 位	14	0	92	1	9	0	118	3
3 位	24	0	77	2	22	0	66	1
4 位	16	0	48	0	54	0	107	6
5 位	24	0	61	1	62	0	127	1
6 位	29	0	67	0	95	0	106	0
7 位	42	0	45	3	136	608	40	577
8 位	39	0	53	2	250	31	22	50
9 位	54	0	47	0	0	0	0	0
10 位	86	0	40	1	0	0	0	0
TOP2	18	23	163	4	20	0	171	4
%	3%	4%	26%	1%	3%	0%	27%	1%

(2)推定内容に関する考察

大分類のみの場合は，すべて答えを見つけることができたが，小分類を含めた場合は少数であるが見つけることが出来なかった場合がある．見つけることが出来なかった分類は，比較対象とした特許群に無かったパターンである．アルゴリズム上，比較対象となる特許群への影響が大きいため，なんらかの対応を検討する必要がある．

また，比較した特許群に含まれる課題分類，手段分類の数と正解の範囲，平均順位との相関関係を表 9 に示す．このうち，最小値を用いて手段の大分類を求めた場合，順位がすべて 1 位であることから範囲がすべて 0 となり，相関係数が求まらないため空欄となっている．

最大値の手法を用いた場合，課題については，値の範囲，順位とも正の相関が見受けられる．最小値を用いた場合についても，やや強い相関がみられる．すなわち，既知の分類数が多いほど，平均順位が下がり，範囲が広がることを意味している．一方，平均値や CF 値は相関が見えにくい傾向があることから，既知の分類数の影響が見えないと予想される．

表 9　特許群における数と範囲・順位の相関係数

手法	範囲	範囲	値の範囲	平均順位
最大	課題	大小	0.67	0.77
	課題	大	0.50	0.77
	手段	大小	0.27	0.29
	手段	大	0.57	0.79
最小	課題	大小	0.63	0.70
	課題	大	0.67	0.46
	手段	大小	0.65	0.57
	手段	大		−0.30
平均	課題	大小	0.62	0.44
	課題	大	0.43	0.13
	手段	大小	0.59	0.36
	手段	大	0.46	0.41
CF 値	課題	大小	0.18	−0.08
	課題	大	−0.19	−0.31
	手段	大小	0.16	−0.15
	手段	大	−0.06	−0.32

(3)今後の課題

今回の実験によって，本手法が活用できる可能性を示せた．しかし，精度はまだ不十分である．そのため精度をあげる方法として，いくつかの観点から検討を行う必要がある．

一つ目としては，類似度そのものの向上である．現在の手法では，得られた形態素をすべて利用している．そのためノイズとなるデータが存在し，それが精度を落としていると考えられる．この点を改善するために，形態素の品詞を固定することを検討する．また現在，課題タグ，手段タグで範囲を決めているが，この点も再度検討する必要がある．範囲が狭い分，先のノイズが大きく影響が出る可能性がある．しかし，範囲を広げるとさらにノイズが加わる可能性も出てくるため，それらを適切に管理する方法が必要となる．その手法として，事前研究[10]において，専門家の協力のもと作成した，特許内の重要と思われる文を抽出するフィルタの利用を検討する．また，事前研究において，N-Gram による類似度が比較的高く得られていることから，形態素の代わりに N-Gram を用いた方法を試みる．また，現在の類似度は文書単位であるため，分類単位の類似度を求めるよう改善する必要がある．これは次の課題とも関係している．

二つ目としては，候補の絞り込み方法である．今回の実験においては，算術平均を用いた手法が比較的良い結果をえている．しかし，まだ不十分であるため，そのほかの候補の絞り込み方法を検討する．現時点では，ナイーブベイズ[11]を援用する手法や幾何平均を利用することを検討している．また，大分類は比較的高いことから，大分類の抽出結果を利用し，大分類と小分類の組み合わせについて抽出する手法を実施する．

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

さらに，今回利用していない F タームや IPC，FI の活用方法についても検討する．これらの情報は直接的に使えなくとも，分類するうえの手がかりになると考える．そのため，F タームや IPC，FI と課題分類，手段分類との相関を求め，その値を用いることを考える．

三つ目としては，改訂手法の評価である．今回用いたデータやその他のデータに対して実験，検証を進める．また，F タームや IPC，FI に対し，本枠組みを適用することでその有用性を評価する．これに対しては，NTCIR のテストコレクション[12][13]の利用を検討している．さらに，本システムの応用として，新規特許と既存特許との関連性の強さを判定することにも試みる予定である．

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．具体的には，研究協力者が利用している課題分類と手段分類を，特許が解決しようとする課題とその手段としては捉え，それらの分類を推定することを試みる．これらの分類は，特許に割り当てられた IPC などとは別に専門家が特許分類に活用しているものである．本研究では，特許文書に付与されているブロックタグに着目し，課題と手段の比較範囲を限定した．本手法では，形態素をベースとした文書ベクトルに特許を変換し，課題，手段についてそれぞれ Cos 類似度で比較する．比較の対象は，新規特許と分類済みの特許群である．特許群は複数あるため，その結果を統合するために，最大値，最小値，平均値，CF 値（確信度係数）を用いた．

本手法の有用性を評価するために，専門家が分類を割り当てた特許 639 件を対象に実証実験を行った．実験においては，一つの特許の分類を別の特許群をもとに推定を行った．その結果得られた，同順数を加えた順位を求め，順位の出現数に基づく評価を行った．結果として，課題の大分類小分類に対して上位 10 位までに含まれる割合は，平均値を用いた時が 21%となった．他の手法が 5%程度であることから，本手法が今回の中では最も良いと評価する．手段の大分類小分類についても同様である．課題の大分類，手段の大分類に対して上位 2 位までに含まれる割合についても同様に，平均値を用いた時が 26%，27%と高くなった．この結果から，平均値を用いるものがもっとも良い結果を得ることができた．しかし，精度としては，ランダムよりも良いものの，まだ不十分である．

今後は，類似度そのものの向上を進める．本手法の根幹となっているのは類似度計算である．よって現在の手法を見直すとともに類似度の計算方法について検討を加える．また，検索範囲によってもその結果が変わるため，その点についても検討が必要である．これらに対しては，N-Gram の活用や，文に対するフィルタの適用，今回利用していない F タームや IPC，FI の活用などを検討している．また，現状では特許文書に対する類似度を基本的に求めているため，課題分類，手段分類に対する類似度を求めるように検討を進める必要がある．さらに，計算時間がかかる場合，実用的ではないため，その点も考慮する．改訂した手法に対しては今回と同様の評価実験を行うとともに，テストコレクションでの評価を行う．

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 24500121）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：特許電子図書館，http://www.ipdl.inpit.go.jp/homepg.ipdl（2014/3/10 アクセス）
[6] 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)
[7] B.G.Buchanan，E.H.Shortliffe：“Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project”， Addison-Wesley(1984). ISBN 978-0-201-10172-0
[8] Lucene-gosen，https;//code.google.com/p/lucene-gosen/（2014/6/13 アクセス）
[9] xdoc2txt，http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html（2014/3/10 アクセス）
[10] 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）
[11] 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)
[12] NTCIR：NTCIR-7 Patent Mining (特許マイニング）テストコレクション，http://research.nii.ac.jp/ntcir/permission/ntcir-7/perm-ja-PATMN.html
[13] NTCIR-8 Patent Mining (特許マイニング テストコレクション)，http://research.nii.ac.jp/ntcir/permission/ntcir-8/perm-ja-PATMN.html

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.