1A3-02

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

A Framework of Estimating Invention Task and Means from Patent Journals Based on the Document Similarity with the Patent Journal Structures

樽松理樹[*1]
Masaki Kurematsu

[*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 from patent journals based on the document similarity. When we use other's invention protected by a patent without permission or violate it, we need a lot of time and load to settle this problem. Therefore it is important to research exists patent journals before submitting own patent or sealing new products. However, it is take long time to check a lot of patents. In order to support this task, I propose a new framework. This framework estimates invention task and means of new patent based on the document similarity between it and each exist patent given invention task and means by experts. This framework divides a patent into 4 parts based on the patent journal structure initially. Next, it gets the similarity of each part and integrates them. Finally, it shows invention task and means ranked in descending order by the similarity in the difference 4 orders. In order to evaluate this system, 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. Therefore, this experimental result shows that is it possible to use this approach to estimate invention task and means from patent journals. I will analyze experimental results, enhance this system based on the result of analysis and evaluate bigger data with experts.

1. はじめに

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」「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[寺岡 10][谷川 13]が提案されてきた．しかし，その多くは，特許電子図書館(IPDL)サービス[工業所有権情報・研修館 94]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．

連絡先：樽松理樹，岩手県立大学，〒020-0611 岩手県滝沢市巣子 152-52，TEL：019-694-2582，FAX：019-694-2501，メール：kure@iwate-pu.ac.jp

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[発明協会 05]．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），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，F タームは審査官が審査に利用する分類記号である．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．

IPC や F ターム，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から課題や手段は推定できると考えられるが，実際には，同じ事柄でも表記が異なる，対象となる範囲が異なるなどの理由から，IPC や F タームのみでの課題や手段の把握は困難である．

2.2 特許の課題と手段

専門家は，特許公報を熟読する前に，その特許が解決しようとする課題の分類と課題を解決するための手段の分類を抽出する．この結果をもとに，権利調査の対象としての重要性を判断

し，重要性の高いものから特許の確認を行う．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，小分類を 62 種設定し，その組み合わせパターン数は，66 である．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，小分類 33 種，組み合わせパターン数 40 であり，一つの大分類における小分類の数は平均 5 種類である．

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．

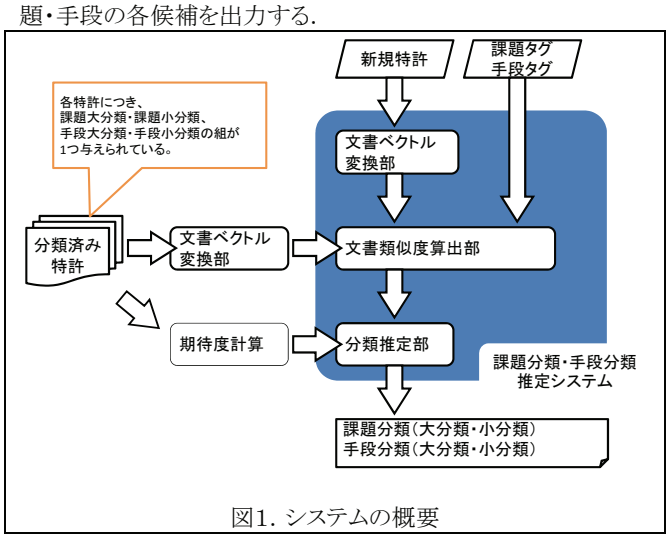

図1. システムの概要

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. 次の方法で，ブロック毎に文から文字列を取り出す．
1.1. ブロックタグが，IPC，FI，出願人の場合は，区切り記号に従い，文字列として取り出す．
1.2. 【発明の名称】【要約】および明細書に含まれるブロックタグの場合，各文から，以下の文字列を切り出す．
①連続する 3 文字の漢字または平仮名，②連続するカタカナ，③連続する英字，④連続する数字，⑤連続する記号文字，⑥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 1.2 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．また，形態素ではなく連続する 3 文字としたのは先行研究の結果[樽松 13a]に基づいている．

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[北 02]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 IPC または FI の場合は，コード類似度 C を式(1)で求める．

$$C = \frac{2\times V_1とV_2の共通コード数}{V_1のコード数+V_2のコード数} \quad \cdots式(1)$$

2. 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(2)で求める．ここで，xt_i は V_1 と V_2 に含まれるパターンiのうちブロックタグが課題タグをみたすものの V_1 での出現数である．また yt_i は V_2 での出現数である．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(3)で求める．ここで，xm_i は V_1 と V_2 に含まれるパターンiのうちブロックタグが手段タグをみたすものの V_1 での出現数である．また，ym_i は V_2 での出現数である．

$$T = \frac{\sum xt_i yt_i}{\sqrt{xt_i^2 \times yt_i^2}} \quad \cdots式(2) \qquad M = \frac{\sum xm_i ym_i}{\sqrt{xm_i^2 \times ym_i^2}} \quad \cdots式(3)$$

3. 上記で求めた値をもとに文書類似度 S を式(4)で求める．各重みは専門家との話し合いで決定している．

$$S = \sqrt{0.3C^2+0.35T^2+0.35M^2} \quad \cdots式(4)$$

3.4 分類推定方法

上記で得られた類似度をもとに，次にあげる方法で課題・手段分類を求める．

(1) 最大類似度…分類済み特許の分類を，類似度の降順に候補として提示する．
(2) 期待度数…分類済み特許の分類を，類似度と分類の組の期待度数の積の降順に候補として提示する．ここで期待度数は式(5)で求める．

課題 i 手段 j の期待度数＝

$$\frac{課題\ i\ の分類済み特許数 \times 手段\ j\ の分類済み特許数}{分類済み特許の総数} \quad \cdots式(5)$$

(3) 平均値…分類済み特許の分類を，分類毎の類似度の平均の降順に候補として提示する．
(4) 個別値…課題の類似度 T と手段の類似度 M についてそれぞれ，(値－最小値)を，(最大値－最小値)で割ることで正規化する．課題分類と手段分類の組ごとに，正規化後の類似度の幾何平均の降順に候補として提示する．(1)から(3)が課題と手段の組で推定しているのに対し，(4)では，それぞれ別々に利用している．

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．実装したシステムのスクリーンショットを図 2 に示す．なお本研究と直接関係ない機能も用意されている．

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 705 件のうち，173 件の特許について，課題と手段の分類の推定を試みる．実験においては，分類を推定する特許以外の特許との文書類似度から分類抽出を行う．また，課題タグとしては“.*課題】”，手段タグとしては“.*手段】”を与える．よって，課題で終わるタグがついているブロックのみを課題の類似度計算に用い，手段で終わるタグがついているブロックのみを手段の類似度計算に用いる．

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．また分類については，「課題大分類・課題小分類・手段大分類・手段小分類」「課題大分類・手段大分類」「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．

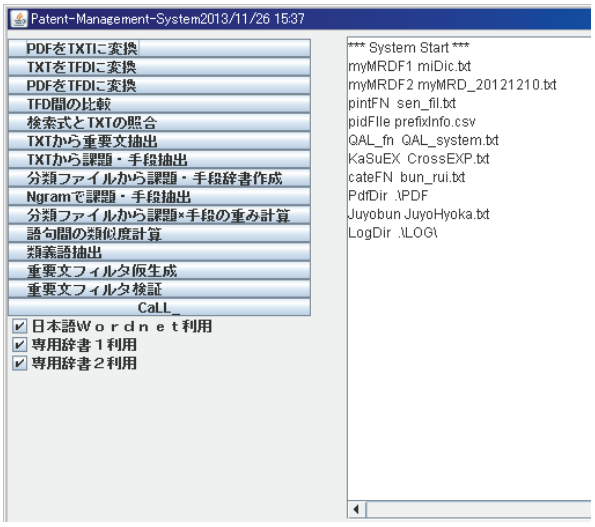

図 2 システムのスクリーンショット

4.2 実験結果

表 1 から表 3 に実験結果を示す．各表において推定手法は，3.4 章であげた分類手法である．課題と手段は対象とした項目を示しており，◎は大分類小分類両方，○は大分類のみである．表 1 は正解の順位の平均値と SD を示し，表 2 は順位の大まかな割合を示す．表 3 は全候補における正解の順位分布(位置)を示しており，たとえば 1～10%は，上位 1～10%以内に正解が出現した特許数を示す．なお，今回推定を試みた特許の分類の中には，類似度を求めるために用いた分類済み特許の中にないものも含まれる．そのため大分類と小分類両方を推定する場合に正解が見つからない場合があった．

表 2 に示すように，順位においては，大分類・小分類をまとめた場合は，最大類似度，個別値，平均，期待度数の順となり，大分類のみでは，最大類似度，平均，個別値，期待度数の順となった．表1の示すように個別値の順位が大きくなっているが，これはパターン数が課題と手段の分類の組み合わせとなり，その数が膨大となるためである．順位が大きいものがあると，その影響を受け平均値が高くなってしまう．一方で，表 3 の順位分布で見た場合は，個別値が最もよく，最大類似度，平均，期待度数の順となる．

表 1. 実験結果(順位の平均値と SD)

課題	手段	順位	最大類似	期待度数	平均	個別値
◎	◎	平均	109.3	215.0	108.0	5393.4
◎	◎	SD	134.5	121.4	81.7	13364.2
◎		平均	109.3	215.0	108.0	5393.4
◎		SD	134.5	121.4	81.7	13364.2
	◎	平均	113.6	216.9	103.9	8002.5
	◎	SD	134.5	121.4	81.7	13364.2
○	○	平均	34.8	54.0	31.0	597.4
○	○	SD	71.7	60.3	46.6	2265.8
○		平均	34.8	54.0	31.0	597.4
○		SD	71.7	60.3	46.6	2265.8
	○	平均	42.5	60.0	36.6	911.0
	○	SD	71.7	60.3	46.6	2265.8

4.3 評価・考察

全体として，大分類のみのほうが高く，手段のほうが課題よりよい結果を得ている．候補となるパターンが大分類だけのほうが大分類・小分類の組み合わせよりも少なく，手段のほうが課題よりも少ないため，このような結果になったと考えられる．ただし，今回の方法においては，ランダムで行う場合よりも精度は高いことから，一定以上の役割は果たしていると考えられる．また，小分類については，専門家間でも意見が分かれるとのことであるため，さらなる検証が必要である．

各手法については，最大類似度，平均，期待度数と個別値とでは分類推定方法が一部異なるので分けて考える．

最大類似度，平均，期待度数については，各表で示したように，最大類似度のものを取り出すパターンが全体的によい結果となっている．このような結果になった理由としては，類似した特許が分類済み特許に含まれていたためと考えられる．一方，過去の例に基づき、逸脱した組み合わせになることを避けるために重みとして導入した期待度数を用いる方法では，むしろ順位を押し下げることとなり，マイナス効果になっている．これは，より多くの可能性を考慮することを目的に導入した平均値でも同様である．これらの理由としては，分類済み特許における分類ごとの特許数のアンバランスが影響を与えていると考えられる．また，これらの手法は，過去の分類を文書類似度の計算とは違う形で利用することなる．そのため，過去の分類が過度に活用されている可能性が高い．そのため，期待度の計算方法を含め，その利用方法を検討する必要がある．

個別値を用いたものは，順位は最大類似度の場合よりも低かったが，また，最大類似度などと異なり，分類済み特許にない課題分類と手段分類の組み合わせを見つけることも可能である．また，順位分布としては，最大類似度よりも上位が多い．これは，

組み合わせのパターン数が多いためである．パターン数は膨大であるが，正解が上位になることから，すべてのパターンではなく，分類候補を上位で絞り込むことが考えられる．また，大分類のみの場合は，他の手法との差が少なくなっているが，同様の傾向がある．以上の結果を踏まえ，個別値を用いる方法が妥当であると考えられる．

また，本研究の先行研究[樽松 13b]においては，分類済み特許に現れる文字列の出現頻度を利用する方法を行った．本手法では，全分類済み特許から今回の文書ベクトルと同様に文字列を切り出し，分類毎の出現頻度を求める．さらに分類に対する TFIDF を求め，それを重みとした．新規特許に対しては，特許中に現れる文字列に対し，その重みの総和を求め，それをもとに分類を推定する．この手法に対し，本手法のほうがより高い精度を得ることができた．また先行研究の手法では，文字列の重みを求めるために膨大な計算が必要なことから，その点からも本手法のほうが有用であると評価できる．

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．今後の課題としては次のことがあげられる．今回推定を行っていない特許に対する推定とその実験結果の検証を行う．検証では，精度ごとに分類し，それぞれの特許の特徴を専門家との議論を行いながら把握する．その過程を通して得る専門家の知識・知見を踏まえ，本手法で利用している各種パラメータの設定方法や手法の再検討，システムの再構築を進める．再構築したシステムに対し，今回と同等の評価実験を専門家と協力しながら実施する．

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，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．

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

[北 02] 北研二 他：情報検索アルゴリズム，共立出版（2002）

[樽松 13a] 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，E-014，pp.211-214（第 2 分冊），（2013）

[樽松 13b] 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）

[谷川 13] 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）

[寺岡 10] 寺岡岳夫：特許情報検索の現状と今後，Japio Year Book 2010，pp.166 - 169（2010）

[工業所有権情報・研修館 94] 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）

[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）

表 2. 実験結果（50 位以内の特許数）

課題	手段	順位	最大類似	期待度数	平均	個別値
◎	◎	1～10 位	30	0	18	22
◎	◎	11～25 位	12	0	8	13
◎	◎	26～50 位	13	11	7	6
◎		1～10 位	30	0	18	22
◎		11～25 位	12	0	8	13
◎		26～50 位	13	11	7	6
	◎	1～10 位	30	0	18	22
	◎	11～25 位	12	0	8	13
	◎	26～50 位	13	11	7	6
○	○	1～10 位	84	23	82	61
○	○	11～25 位	39	46	34	21
○	○	26～50 位	24	46	23	13
○		1～10 位	84	23	82	61
○		11～25 位	39	46	34	21
○		26～50 位	24	46	23	13
	○	1～10 位	84	23	82	61
	○	11～25 位	39	46	34	21
	○	26～50 位	24	46	23	13

表 3. 実験結果（上位 50%以内の特許数）

課題	手段	順位分布	最大類似	期待度数	平均	個別値
◎	◎	1～10%	124	79	94	168
◎	◎	11～25%	20	26	14	5
◎	◎	26～50%	19	47	42	0
◎		1～10%	124	79	94	168
◎		11～25%	20	26	14	5
◎		26～50%	19	47	42	0
	◎	1～10%	124	79	94	168
	◎	11～25%	20	26	14	5
	◎	26～50%	19	47	42	0
○	○	1～10%	150	127	124	173
○	○	11～25%	14	33	33	0
○	○	26～50%	5	12	11	0
○		1～10%	150	127	124	173
○		11～25%	14	33	33	0
○		26～50%	5	12	11	0
	○	1～10%	150	127	124	173
	○	11～25%	14	33	33	0
	○	26～50%	5	12	11	0