2K3-5

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

A study of estimation issues and methods from patent documents based on term frequency

榑松理樹[*1]
Masaki KUREMATSU

[*1] 岩手県立大学
Iwate Prefectural University

In this paper, I proposed a framework of estimating invention task and means based on term frequency in patent journal. 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 novel framework which estimates invention task and means of new patent based on term frequency in patent journal. First, this framework extracts terms from abstracts of patents which experts identified invention task and means in advance. Secondly, it calculates the term weight based on term frequency. Thirdly, it converts an unclassified patent to a document vector using by term weights. Finally, it shows invention task and means ranked in descending order by the similarity among document vectors. 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 an expert. This experimental result shows it is possible to use this approach to estimate invention task and means from patent journals. I will analyze experimental results and enhance this system based on the result of analysis.

1. はじめに

特許公報[発明協会 05]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」等の問題から作業負荷が大きい．このような作業負荷を軽減し，特許公報活用の有効性，効率性を向上させるために，これまでにいくつかのコンピュータによる支援方法[寺岡 10] [谷川 13] [藤井 12]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 [研修館]に代表されるような検索システムである．これらのシステムでは，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[榑松 14][榑松 15]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公報を，2.2 で述べる特許公報に付与された【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】などのブロックタグで分割し，分割された範囲ごと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．本手法においては一定の推定精度をあげることが出来た．しかし，一つの特許公報との類似度やブロックタグによる偏りなどの影響から，精度は不十分である．

以上の背景から，本稿では次のような修正を加えた手法を提案する．一つ目の修正としては，課題・手段分類のために用いる語句の抽出範囲を，従来の特許公報のブロックごとから，特許公報の要約に変更する．これは，他のブロックから抽出されるノイズとなる語を軽減することと，計算量の削減が目的である．二つ目としては，出現回数のカウント方法を，従来のブロック単位ではなく，特許公報全体とする．これは，ブロック単位で行った場合に値の差が小さくなる傾向があったため，それを押さえ，識別率を挙げることが目的である．三つ目として，抽出を試みる課題や手段の分類を出現数に基づき統合する．これにより候補が絞り込め，精度があがることが期待できる．以上の変更を行ったシステムを実装し，実際に評価を行った．

連絡先：020-0693 岩手県滝沢市巣子 152-52 岩手県立大学 ソフトウェア情報学部，電話 019-694-2582，FAX：019-694-2501，eMail：kure@iwate-pu.ac.jp

2. 語句出現頻度を利用した特許公報からの課題・手段推定システム

2.1 システム概要

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，分類済み特許公報，対象となる新規特許をそれぞれ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，手段の候補を推定する．以降で特許公報について説明したあと，各部分の説明を加える．

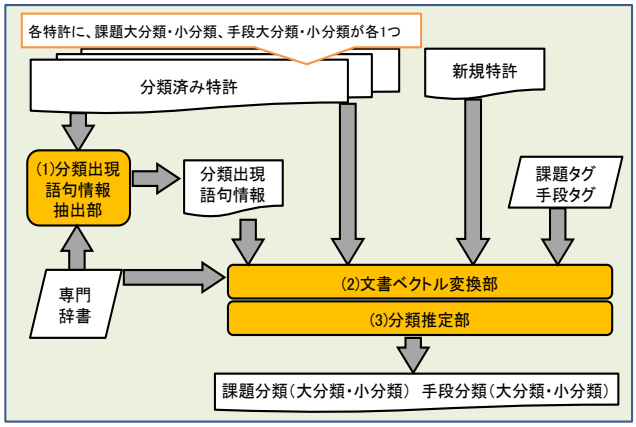

図1：システムの概要

2.2 対象とする特許公報

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），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[藤井 12] [研修館]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．

本システムでは，専門家によって一定の範囲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と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つから構成される．また，大分類ごとに小分類はことなる．

2.3 分類出現語句情報抽出部

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

(1) 分類の絞込み･･･分類出現語句情報を抽出する特許公報に付与された分類ラベルのうち，その出現割合が与えられた閾値以下の場合は統合する．統合する場合は，それ専用のラベルを用いる．

(2) 対象とする文章の抽出…特許公報に含まれる要約文から課題について述べている課題文，手段について述べている手段文にブロックタグを用いて抽出する．

(3) 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，(d)専門辞書中の代表語のいずれかの方法で語句を抽出する．(a)の形態素列としては，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．ただし，連続する語は 2 語までとする．(b)(c)のカタカナ列，英字列はそれぞれ連続する1文字以上のカタカナ，英字の並びである．(d)の代表語は，専門家によって構築された専門辞書において，ある語句の概念を示す代表的な語句として定義されたものである．特許公報中に語句が出現した場合，その語句とともに代表語も抽出する．

(4) 重さの決定…(3)で抽出した語句について，分類推定用の重みを，式(1)，式(2)を用いて求める．式(1)は語句の出現回数，式(2)は語句の出現文書数に着目している．各式で t_i は語句，C_j は分類を，t_*，C_*はそれぞれ任意の語句，分類を示す．式(1)において， $tf(t_i,C_j)$は分類 C_j における語句 t_i の出現回数を， $tf(t_i,C_*)$は語句 t_i の出現総数を示す．式(2)において， D は全文書数，$df(t_i,C_j)$は分類 C_j における語句 t_i の出現文書数， $df(t_i,C_*)$は語句 t_i の出現文書総数，$df(t_*,C_j)$は分類 C_j の出現文書総数をそれぞれ示す．なお，式(2)は相互情報量の考えを援用している．

$$wt1(t_i,C_j)=\frac{tf(t_i,C_j)}{tf(t_i,C_*)} \quad \text{式(1)}$$

$$wt2(t_i,C_j)=\log\left(\frac{df(t_i,C_j)/D}{df(t_i,C_*)/D\times df(t_*,C_j)/D}\right) \quad \text{式(2)}$$

以上で求めた分類毎の語句とその重さの組を分類出現語句情報とする．

2.4 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，各特許を次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．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と同じ方法で語句を取り出す．それらと分類出現語句情報とを照合し，分類毎の重さ $V(C_j)$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，式(3)を用いて求める．式(3)において，$wt_*(t_i,C_j)$は，$wt_1(t_i,C_j)$ または $wt_2(t_i,C_j)$を意味し，$tf(t_i,d_k)$は文書 d_k における語句 t_i の出現回数を示す．

$$V(C_j)=\sum\log(1+wt_*(t_i,C_j)\times tf(t_i,C_j)) \quad \text{式(3)}$$

2.5 分類推定部

分類推定部では，文書ベクトルをもとに分類を推定する．

一つ目の方法としては，文書ベクトルの各値を比較し，最大値を持つ要素，すなわち分類を推定結果として出力する．

二つ目の手法としては，文書ベクトル間の類似度を求め，K-NN 法によって分類を推定する．類似度としては，カイ二乗（以後、χ^2），Cos 類似度，ベクトル間の距離［北 02］を用いる，それぞれの計算式を，式(4)から式(6)に示す．ここで，なお上位 K 個のうち最も出現数が多い分類が複数ある場合は，より上位にある分類を候補として抽出する．これらの式において，V は推定を行う特許の課題または手段の文書ベクトル，Wi 分類済み特許の課題または手段のi番目の文書ベクトルを示し，v_j，$w_{i,j}$は，それぞれ V, Wi の j 番目の要素の値を示す．

$$sim(V,W_i)=\sum\frac{(v_j-w_{i,j})^2}{w_{i,j}} \quad \text{式(4)}$$

$$sim(V,W_i)=\frac{\sum(v_j)(w_{i,j})}{\sqrt{\sum(v_j)^2}\sqrt{\sum(w_{i,j})^2}} \quad \text{式(5)}$$

$$sim(V,W_i)=\sqrt{\sum(v_j-w_{i,j})^2} \quad \text{式(6)}$$

3. 評価実験

3.1 実験概要

提案手法の有用性を評価するために，2 章で示した考えをもとに JAVA にて実装したシステムを用いて，以下の条件のものと評価実験を行った．なお形態素解析としては，lucene-gosen-4.0.0-naist-chasen [Lucene-gosen]を用いた．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公報のうち，1998 年から 2008 年までの 283 件から分類出現語句情報を抽出し，2009 年から 2010 年の 59 件の課題分類と手段分類を推定する．

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 K-NN 法の K の値としては，5 を用いた．

評価としては，専門家が付けた分類を正解とし，それが何番目に推定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」「課題大分類のみ」「手段大分類・手段小分類」「手段大分類のみ」それぞれについて評価する．

3.2 分類出現語句情報結果

分類出現語句情報の抽出結果を表1に示す．2.3 で述べた閾値としては，なし，5%以下は統合，10%以下は統合の 3 パターンを用意した．表1で示したように，5%以下を統合することで，大分類小分類の両方の場合は，課題で約 7%、手段で約 21%にまで統合された．このことから分類ごとに出現数に偏りが大きいといえる．

3.3 課題分類・手段分類推定結果

分類の推定結果を表 2 から表 5 に示す．各表において，組の“大小”は大分類ラベルと小分類ラベルの組を，“大－”は大分類のラベルのみを推定することを意味する．閾値は，3.2 で述べた閾値と同じである．“wt#”は 2.3 で述べた重みの種類を示す．手法にあげた“χ^2”，“Cos”，“距離”はそれぞれ類似度の計算方法を示し，“べ1”“べ 2”“べ 3”は，それぞれ 2.4 で述べた手法で生成した文書ベクトルにおいて，値が 1 位，2 位，3 位に正解が出現した割合を示す．また太字になっているものは，推定する分類ラベルおよび利用する分類出現語句情報が同一の範囲での最大値を示す．

3.4 評価・考察

一部を除けばランダムで選択するよりも高い値となった．このことから，提案手法は，精度は低いが有用に働く可能性がある．

重みに関しては，閾値が無い場合を除けば，wt1 のほうが良い傾向が見られた．これは wt2 が文書数に着目しており，統合によって文書数の差が減ったためと思われる．

類似度計算手法は，課題については，候補が多い場合は Cos 類似度がよいが，候補が少ない場合は距離の方が若干上回っている．一方手段は，Cos 類似度が高い．これはラベル毎に抽出した語句の傾向の影響が出ていると考えられる．また，課題においては，文書ベクトル単独での推定が，複数文書ベクトルからの推定を上回る場合が数件見られたが，全体としては複数文書ベクトルの比較の方が精度は高い．これは特許公報全体の比較となるためと考えられる．

本提案手法によって一定の課題・手段の推定は行えた場，その精度はまだ不十分である．今後の課題としては，課題・手段ラベル数との比較など，今回の実験結果の分析があげられる．さらにそれらの分析に基づく新たな推定手法の考案と検証を行う．あたらな推定手法としては，潜在意味解析［佐藤 15］の援用やラフセット理論［森 13］に基づく新たな手法の構築を試みる．潜在意味解析の利用においては，本題材では文書単位の解答がすでに存在することとなることから，その点を考慮することが必要となる．また，分類出現語句の適切な統合方法についても検討を進める．

表1：分類出現語句抽出結果

		課題		手段	
閾値	区分	ラベル	語彙	ラベル	語彙
なし	大小	55	7684	33	12189
	大－	12	9782	8	7724
5%以上	大小	4	2932	7	8004
	大－	9	9401	6	7592
10%以上	大小	1	2298	4	6772
	大－	4	5958	5	7358

4. 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，特許公報で述べられている，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，専門家により事前に分類された特許公報の要約文における語句の出現情報をもとに未分類特許公報の分類を推定する．専門家の協力のもとに行った評価実験においては，ランダムで選択した精度よりも高い値となったが，ラベル数の偏りに影響が大きい結果となった．今後は，実験結果の分析に基づく推定方法の改善，処理結果の反映による精度の向上などによる改善を進める．

表 2:課題に対する推定結果

組	重み	類似度	閾値		
			なし	5%以上	10%以上
大小	wt1	χ2	5.1%	**88.1%**	**100.0%**
大小	wt1	Cos	**16.9%**	84.7%	**100.0%**
大小	wt1	距離	3.4%	**88.1%**	**100.0%**
大小	wt1	ベ1	**16.9%**	**88.1%**	**100.0%**
大小	wt1	ベ2	6.8%	8.5%	0.0%
大小	wt1	ベ3	0.0%	1.7%	0.0%
大ー	wt1	χ2	28.8%	30.5%	30.5%
大ー	wt1	Cos	**33.9%**	**32.2%**	35.6%
大ー	wt1	距離	11.9%	11.9%	11.9%
大ー	wt1	ベ1	18.6%	18.6%	**47.5%**
大ー	wt1	ベ2	13.6%	13.6%	30.5%
大ー	wt1	ベ3	8.5%	8.5%	6.8%

表 3:課題に対する推定結果

組	重み	類似度	閾値		
			なし	5%以上	10%以上
大小	**wt2**	χ2	**6.8%**	**81.4%**	**100.0%**
大小	**wt2**	Cos	**16.9%**	**79.7%**	**100.0%**
大小	**wt2**	距離	**3.4%**	**86.4%**	**100.0%**
大小	**wt2**	ベ1	**13.6%**	**32.2%**	**100.0%**
大小	**wt2**	ベ2	**8.5%**	**42.4%**	**0.0%**
大小	**wt2**	ベ3	**1.7%**	**20.3%**	**0.0%**
大ー	**wt2**	χ2	**16.9%**	**28.8%**	**27.1%**
大ー	**wt2**	Cos	**30.5%**	**27.1%**	**30.5%**
大ー	**wt2**	距離	**0.0%**	**30.5%**	**25.4%**
大ー	**wt2**	ベ1	**3.4%**	**16.9%**	**30.5%**
大ー	**wt2**	ベ2	**13.6%**	**8.5%**	**15.3%**
大ー	**wt2**	ベ3	**6.8%**	**18.6%**	**42.4%**

表 4:手段に対する推定結果

組	重み	類似度	閾値		
			なし	5%以上	10%以上
大小	wt1	χ2	8.5%	20.3%	18.6%
大小	wt1	Cos	**33.9%**	45.8%	**69.5%**
大小	wt1	距離	23.7%	40.7%	61.0%
大小	wt1	ベ1	30.5%	**50.8%**	62.7%
大小	wt1	ベ2	6.8%	22.0%	28.8%
大小	wt1	ベ3	8.5%	6.8%	6.8%
大ー	wt1	χ2	13.6%	20.3%	20.3%
大ー	wt1	Cos	**57.6%**	**57.6%**	**57.6%**
大ー	wt1	距離	23.7%	23.7%	23.7%
大ー	wt1	ベ1	40.7%	40.7%	42.4%
大ー	wt1	ベ2	18.6%	18.6%	22.0%
大ー	wt1	ベ3	13.6%	13.6%	13.6%

表 5:手段に対する推定結果

組	重み	類似度	閾値		
			なし	5%以上	10%以上
大小	wt2	χ2	**11.9%**	**11.9%**	**6.8%**
大小	**wt2**	Cos	**32.2%**	**44.1%**	**67.8%**
大小	**wt2**	距離	**6.8%**	**25.4%**	**20.3%**
大小	**wt2**	ベ1	**18.6%**	**39.0%**	**40.7%**
大小	**wt2**	ベ2	**16.9%**	**5.1%**	**42.4%**
大小	**wt2**	ベ3	**8.5%**	**18.6%**	**16.9%**
大ー	**wt2**	χ2	**10.2%**	**8.5%**	**10.2%**
大ー	**wt2**	Cos	**55.9%**	**55.9%**	**55.9%**
大ー	**wt2**	距離	**23.7%**	**23.7%**	**23.7%**
大ー	**wt2**	ベ1	**54.2%**	**50.8%**	**52.5%**
大ー	**wt2**	ベ2	**15.3%**	**20.3%**	**20.3%**
大ー	**wt2**	ベ3	**13.6%**	**11.9%**	**10.2%**

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

[藤井 12] 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)

[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)

[研修館] 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 (2016/3/20 アクセス)

[北 02] 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)

[樽松 14] 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, 人工知能学会全国大会第 26 回(2014)

[樽松 15] 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, 人工知能学会全国大会第 27 回(2015)

[Lucene-gosen] Lucene-gosen , https://github.com/lucene-gosen/lucene-gosen (2016/3/20 アクセス)

[森 13]森典彦, 森田小百合:“人の考え方に最も近いデータ解析法―ラフ集合が意思決定を支援する”, 海文堂出版 (2013)

[Pedro 97] 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)

[佐藤 15]佐藤一誠:“トピックモデルによる統計的潜在意味解析”, コロナ社(2015)

[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)