機械学習を用いた特許文書分析方法の提案と評価

M2018SE010　三浦　敦子
指導教員　青山　幹雄

1 研究の背景

近年，機械学習を用いた特許文書の分析方法が提案されている．特許文書の分析では，分析目的（先行技術調査，技術動向調査など）[5]，特許文書の特殊性（造語，複合語が多い，特許独特の表現など）を考慮する必要がある．

本稿では先行技術調査に焦点を当て，特許文書の特殊性に着目し，機械学習を用いた特許文書分析方法を提案する．

2 研究の課題

本稿では，以下の 3 点を研究課題とする．

RQ1: 再現率を向上しつつ，特許調査の効率化は可能か
RQ2: 特許文書分類（機械学習）の処理効率向上は可能か
RQ3: 特許文書分析プロセスにおいて提案方法は有効か

3 関連研究

3.1 機械学習による新規性キーワード抽出方法

特許文書の解析について，請求項の構造解析を利用して，次の 2 つの仮説に基づき請求項から新規性に関するキーワードを抽出する方法が提案されている[6]．

(1) 独立請求項において新規性および進歩性がある箇所は他の構成要素にあまり限定されない．
(2) 独立請求項において従属請求項により限定される箇所は出願人が新規性および進歩性があると考える箇所である．

さらに，この新規性キーワード抽出方法を評価する正解データを自動作成するアプローチが提案されている．

3.2 機械学習による高類似文書抽出方法

特許文書の先行技術調査について，類似度の高い文書を抽出する方法が検討されている[1]．この高類似文書抽出方法における Doc2Vec[4]を用いた文書のベクトル化処理の概要を図 1 に示す．

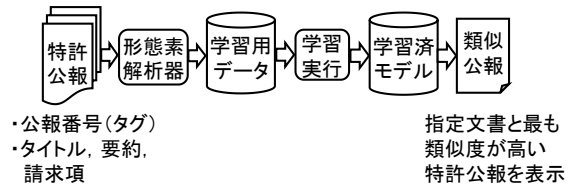

図 1　高類似文書抽出方法

4 アプローチ

本稿の着眼点とアプローチを図 2 に示す．本稿では，特許文書（先行技術文献）の探索範囲から明細書を除外せず，かつ特許分類[3]を使用せずに探索範囲を限定する方法として，明細書の探索範囲を発明のポイントとなる可能性が高い箇所に限定するアプローチをとる．

着眼点 1 は，特許文書の探索範囲から明細書を省くと探索範囲が狭過ぎる点である．審査では，明細書の記載に基づき拒絶されることが多く，例 2 のように探索範囲から明細書を除外すると，探索範囲が狭過ぎると考えられる．

着眼点 2 は，特許分類を用いて特許文書数を絞ることが考えられるが，特許分類を調べて把握することは特に特許調査に慣れていない者にとっては手間がかかる作業であり，また，特許分類を適切に選択できないと特許文書数を絞り過ぎてしまう可能性がある点である．

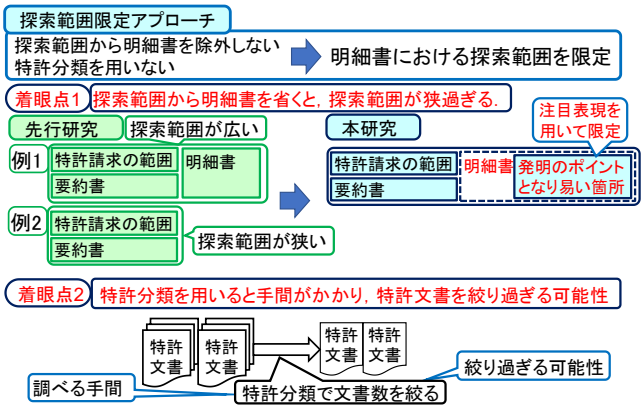

図 2　探索範囲限定アプローチ

5 提案方法

5.1 探索範囲の限定

本稿の特許文書分析プロセスでは，特許文書（先行技術文献）の分類（機械学習）の前に，注目表現を用いて文書フィルタにより，探索範囲を発明のポイントとなる可能性が高い箇所に限定する方法を提案する（図 3）．

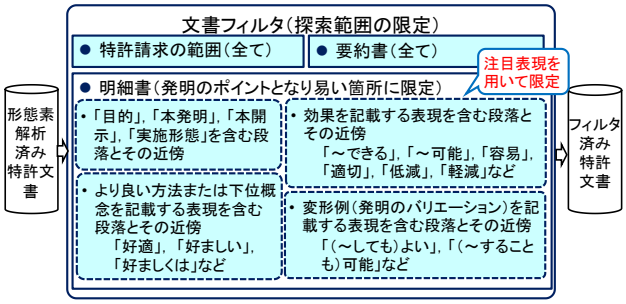

図 3　文書フィルタ（探索範囲の限定）

特許請求の範囲と要約書には発明のポイントとなる内容が記載されるため，全てを探索範囲とする．明細書については，注目表現を用いて探索範囲を限定する．明細書には，主に発明の内容の説明が記載され，関連する様々な情報も記載されるので，特許請求の範囲よりも文字数が多い．したがって，明細書については，発明の内容を含む箇所および発明の内容との関連性が高い箇所（発明のポイントとなる箇所）に探索範囲を限定することで，調査対象の請求項との比較がより適切になると考えられる．

5.2 注目条件の設定

明細書における注目表現の設定について説明する．明細書には，多くの場合，背景技術（先行技術），発明の概要，および発明を実施するための形態（発明を具体的にどのように実施するかなど）が記載される（図 4）．

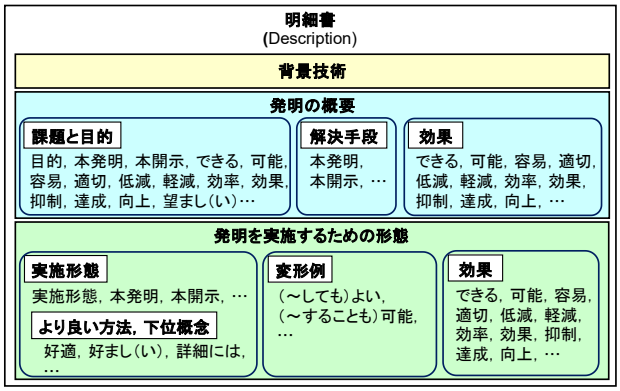

図 4　明細書の構成

5.2.1 背景技術

背景技術には，従来技術(先行技術)が記載されるので，背景技術を対象とした注目表現を設定する必要はない．

5.2.2 発明の概要

発明の概要には，多くの場合，課題と目的，解決手段，および効果が記載される．

5.2.2.1 課題と目的

課題には従来技術における問題が記載されることもあれば，どのようになることが望ましいか(望まれる結果)が記載されることもある．発明の目的を記載する際には，「本発明の目的は…」，「本開示の目的は…」といった表現が用いられることが多いので，「本発明」，「本開示」，「目的」を注目表現として用いる．望まれる結果や目的としては，効果や改善事項が記載されることが多く，次のような表現を注目表現として用いる．

「できる」，「可能」，「容易」，「適切」，「低減」，「軽減」，「効率」，「効果」，「抑制」，「達成」，「向上」，「望まし（い）」

5.2.2.2 解決手段

解決手段には課題を解決する手段(目的を達成する手段)が記載され，発明の内容と補足事項が記載されることが多いので，発明の内容を記載する際に用いられることが多い「本発明」，「本開示」を注目表現として用いる．

5.2.2.3 効果

効果には課題を解決する手段(発明)によって得られる効果や改善される事項が記載される．したがって，望まれる結果や目的の場合と同様の表現を注目表現として用いる．

5.2.3 発明を実施するための形態

発明を実施するための形態には，多くの場合，実施形態，変形例，および効果が記載される．

5.2.3.1 実施形態

実施形態には発明を具体的にどのように実施するかが記載される．実施形態を記載する際，「実施形態」という表現を用いることが多い．また，発明に関連する内容を記載する際に，「本発明」，「本開示」といった表現もある．したがって，「実施形態」，「本発明」，「本開示」を注目表現として用いる．

実施形態における説明において，発明のポイントとなる内容について，上位概念，中位概念，下位概念に分けて段階的に内容を狭めて記載することがある．下位概念がより良い方法である場合もある．下位概念やより良い方法を記載する際には，「好ましくは,…より好ましくは，…」，「好適には，…」，「…がすることが好適である」，「詳細には，…」といった表現を用いることが多い．よって，「好まし」，「好適」，「詳細には」を注目表現として用いる．

5.2.3.2 変形例

変形例としては，発明のバリエーションが記載され，「…してもよい」，「…することも可能」といった表現を用いることが多いので，「よい」，「可能」を注目表現として用いる．

5.2.3.3 効果

発明を実施するための形態における効果としては，実施形態や変形例の効果が記載される．効果に関連する内容(効果の近くに記載されている内容)は発明のポイントとなる可能性が高い．よって，効果を記載する際に用いる表現を注目表現とする．具体的には，上記の発明の概要の効果の場合と同様の表現を注目表現として用いる．

5.3 機械学習による特許文書分析プロセス

特許文書分析プロセスを図 5 に示す．

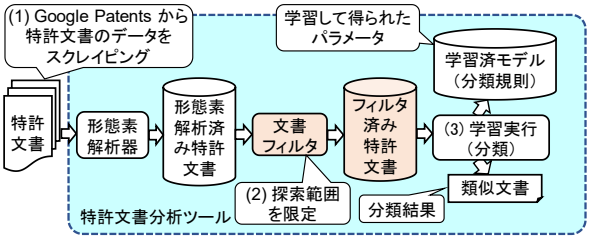

図 5　特許文書分析プロセス

(1) Google Patents から特許文書のデータをスクレイピングし，形態素解析を行う．

(2) 形態素解析済みの特許文書を，上述のアプローチによる文書フィルタを用いてフィルタリングする．

(3) フィルタ済みの特許文書に対して Doc2Vec を用いた機械学習を実行し，調査対象となる請求項との類似度に応じて特許文書を分類する．

機械学習の結果，学習モデル(分類規則)と分類結果として類似文書が得られる．

5.4 類似度分析方法

調査対象の請求項と各特許文書(先行技術文献)との類似度分析方法を図 6 に示す．

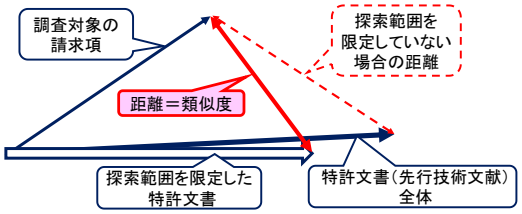

図 6　類似度分析方法

本稿では，類似度の分析に Doc2Vec を用いる．Doc2Vec は，任意の長さのテキスト(単語の集合)をベクトル化する技術であり，テキストの意味を学習できる．各特許文書について，調査対象の請求項のベクトルと探索範囲を限定した特許文書のベクトルとの距離(内積)を類似度として算出し，特許文書を分類する．

一般に，異なる文書で，同じ内容の指摘に異なる表現が用いられることがある．また，特許文書では，造語，複合語などがしばしば用いられる．Doc2Vec を用いることで，異なる文

書で同じ内容の指摘に異なる表現が用いられている場合でも，適切に類似度を分析できる可能性がある．

6 例題への適用

6.1 適用目的

提案方法を特許出願の先行技術調査（新規性調査）に適用して，調査対象の請求項と先行技術となる各特許文書の類似度を分析することで，提案方法の有効性を評価する．

6.2 適用対象

提案方法を適用する調査対象の請求項は，次に示す「特許検索競技大会 過去問 2016 アドバンストコース 電気分野」の問 2 の請求項 1 とする[2].

> 寝具に脈拍あるいは心拍を検知するセンサーを備え、就寝者の脈拍あるいは心拍を前記センサーが検知して、入眠状態であるか否かを判定し，判定結果により湿度と温度をコントロールすることを特徴とする睡眠環境調装置。

提案方法の適用対象は，Google Patents が提供する特許文献とする．提供される全ての特許文献を適用対象とするとデータ量が膨大となるため，公開時期等でデータ量を絞る．

試行として，表 1 に示す条件で特許文書の絞り込みを行い，次の 2 種類の方法でスクレイピングを行った．

(1) 明細書のみ抽出

(2) 明細書，特許請求の範囲，要約書を抽出

表 1　特許文書の絞り込み条件

公開日	2003 年 1 月 1 日から 2016 年 12 月 31 日まで
出願国	日本
言語	日本語
検索語	(脈 OR 心) AND 温 AND 湿

明細書のみ抽出した特許文書と明細書，特許請求の範囲，要約書を抽出した特許文書のそれぞれに対して、形態素解析後 Doc2Vec を用いた機械学習を実行し，類似度が高い特許文書上位 10 件を抽出した．この試行では，文書フィルタによる探索範囲の限定は行っていない．

6.3 適用結果

6.3.1 明細書のみ抽出した特許文書

明細書のみ抽出した特許文書に対して機械学習を行った結果類似度が高い特許文書上位 10 件の公報番号と類似度を表 2 に示す．類似度は調査対象の請求項と特許文書（先行技術文献）をそれぞれベクトル化し，ベクトルの内積を算出することで得る．類似度の値が 1 に近い程，類似度が高い．

表 2　明細書のみ抽出した特許文書の学習結果

No.	順位	公報番号	類似度
1-1	1	JP-5392906-B2	0.799
1-2	2	JP-3465654-B2	0.766
1-3	3	JP-2015054224-A	0.701
1-4	4	JP-3455631-B2	0.688
1-5	5	JP-3514919-B2	0.684
1-6	6	JP-3465849-B2	0.677
1-7	7	JP-3448807-B2	0.677
1-8	8	JP-2014235783-A	0.661
1-9	9	JP-3430956-B2	0.656
1-10	10	JP-3421008-B2	0.649

6.3.2 明細書，特許請求の範囲，要約書を抽出した特許文書

明細書，特許請求の範囲，要約書を抽出した特許文書に対して機械学習を行った結果，類似度が高い特許文書上位 10 件の公報番号を類似度と共に表 3 に示す．

表 3　明細書，特許請求の範囲，要約書を抽出した特許文書の学習結果

No.	順位	公報番号	類似度
2-1	1	JP-3380314-B2	0.828
2-2	2	JP-2003502377-A	0.817
2-3	3	JP-3809972-B2	0.771
2-4	4	JP-2005115121-A	0.763
2-5	5	JP-3362203-B2	0.762
2-6	6	JP-2009508957-A	0.758
2-7	7	JP-3478963-B2	0.755
2-8	8	JP-2003294725-A	0.747
2-9	9	JP-2003295388-A	0.743
2-10	10	JP-2006523675-A	0.738

7 評価と考察

7.1 例題への適用結果の評価

7.1.1 明細書のみ抽出した特許文書

調査対象の請求項におけるキーワードの上位 10 件の特許文書における出現有無を表 4 と表 5 に示す．

調査対象の請求項において「睡眠環境調整装置」は「ことを特徴とする」の前に記載された処理を行う装置の総称として用いられている．したがって，「環境」と「調整」は特許文書（先行技術文献）に記載されている必要はないが，学習時にはこれらの語も含めてベクトル化されるため，キーワードに含めた．

表 4 と表 5 において，キーワードが特許文書に含まれている場合は○とし，含まれていない場合は×とした．特許文書にキーワードそのものは含まれていないがキーワードと意味が近い語が含まれている場合には△とした．

表 4　調査対象請求項におけるキーワードの出現有無

キーワード	センサ（ー）	コントロール	環境	調整
1-1	○	×	○	×
1-2	○	△	×	△
1-3	○	△	○	△
1-4	○	△	○	○
1-5	○	△	○	○
1-6	○	○	×	○
1-7	○	△	×	×
1-8	○	○	○	○
1-9	×	×	×	×
1-10	○	×	×	×

表 5　調査対象請求項におけるキーワードの出現有無

キーワード	寝具	脈拍	心拍	就寝	入眠	湿度	温度
1-1	×	○	○	△	△	○	○
1-2	×	×	×	×	×	×	○
1-3	×	○	○	×	×	○	○
1-4	×	×	×	×	×	○	○
1-5	×	×	×	×	×	○	○
1-6	×	×	○	×	×	○	○
1-7	×	×	×	×	×	○	○
1-8	○	×	○	○	○	△	○
1-9	×	×	×	×	×	△	△
1-10	×	×	×	×	×	△	△

1 位（No. 1-1）の特許文書には，センサにより脈拍を検知し，学習者の眠気を推定することが記載されているが，調査対象の請求項と異なり，脈拍を検知することで入眠状態であるか否かまで判定することや，判定結果に基づいて湿度と温度をコントロールすることは記載されていない．したがって，1 位（No. 1-1）の特許文書は調査対象の請求項の新規性を否定するものとは考えられない．

装置の何らかの制御についての発明では特に「コントロー

ル」,「環境」,「調整」は一般的に用いられることが多く，これらの語は請求項の内容を特徴付けるものではない．また，「センサ」は何らかの情報を検知する発明では用いられることが多く，請求項の内容を特徴付けるものではない．上述のとおり，「環境」と「調整」は特許文書（先行技術文献）に記載されている必要はない．上位の特許文書において「コントロール」,「環境」,「調整」,「センサ」が出現しているものが多いことから，これらの語が特許文書の分析（先行技術調査）において有効でなく，悪影響を及ぼしている可能性があると考えられる．

7.1.2 明細書，特許請求の範囲，要約書を抽出した特許文書

調査対象の請求項におけるキーワードの上位 10 件の特許文書における出現有無を表 6 と表 7 に示す．

表 6　調査対象請求項におけるキーワードの出現有無

キーワード	センサ（一）	コントロール	環境	調整
2-1	○	△	○	○
2-2	×	○	×	×
2-3	×	○	○	○
2-4	×	○	○	○
2-5	○	○	×	○
2-6	×	△	○	○
2-7	×	△	○	○
2-8	○	△	×	○
2-9	×	○	○	○
2-10	×	△	○	△

表 7　調査対象請求項におけるキーワードの出現有無

キーワード	寝具	脈拍	心拍	就寝	入眠	湿度	温度
2-1	×	×	×	×	×	○	○
2-2	△	×	×	△	×	×	○
2-3	×	×	×	×	×	○	△
2-4	×	×	×	×	×	○	○
2-5	×	×	×	×	×	△	△
2-6	×	×	×	△	△	×	○
2-7	×	×	×	×	×	△	○
2-8	×	×	×	×	×	×	○
2-9	×	×	×	×	×	○	○
2-10	×	×	×	×	×	×	○

1 位（No. 1-1）の特許文書は，生物体の育成環境内は水分子付加負イオン含有空気雰囲気を形成し，生物体の育成の適度な温湿度環境を保持して，育成環境内の菌繁殖を発生源において抑制する菌抑制方法に関し，調査対象の請求項の新規性を否定するものとは考えられない．

明細書のみ抽出した特許文書の場合と同様，上位の特許文書において，「コントロール」,「環境」,「調整」が出現しているものが多いことから，これらの語が悪影響を与えている可能性があると考えられる．

また，明細書のみ抽出した特許文書の場合と比べて，「寝具」,「脈拍」,「心拍」,「就寝」,「入眠」,「湿度」,「温度」の出現率が低い．特許請求の範囲と要約書には発明のポイントとなる内容が記載されるが，これらを含めた場合の方が，明細書のみ抽出した特許文書の場合よりも，調査対象の請求項との類似性が低い文書が類似する特許文書として抽出されている．探索範囲が広がったことで，類似度分析がより困難になった可能性があると考えられる．

7.2 考察

適用結果から，調査対象の請求項には，発明の内容を特徴付けるものではない語も含まれており，これらの語が特許文書の分析（先行技術調査）において有効でなく，むしろ悪影響を与えている可能性があることが明らかになった．例えば、調査対象の請求項は「…睡眠環境調整装置。」と記載されており，請求項の最後に記載される名詞句「睡眠環境調整装置」は，発明の内容を示すというよりはむしろ，発明の主題（「装置」であるか「方法」であるかなど）を示す．

また，明細書のみ抽出した特許文書の場合よりも，明細書，特許請求の範囲，要約書を抽出した特許文書の場合の方が，調査対象の請求項との類似性が低い文書が抽出されている．このことから，探索範囲が広がると類似度分析がより困難になる可能性があると言える．

調査対象の請求項については，請求項に記載の発明の内容を特徴付けない語の影響を低くするようにした方が良いと考えられる．したがって，機械学習において請求項に記載の語句のうち，様々な分野で一般的によく用いられる語句や請求項の最後に記載される名詞句を省いたり，これらの語句の重みを低くしたりする方法が有効であると推定できる．

特許文書の探索範囲については，上記の探索範囲制限方法を用いることが有効であると考えられる．明細書，特許請求の範囲，要約書を探索範囲とするが，明細書については注目表現を用いて文書フィルタにより探索範囲を発明のポイントとなる可能性が高い箇所に限定する．

また，特許文書分析において TF-IDF (Term Frequency-Inverse Document Frequency) を用いて，特定の文書に多く出現する語に重み付けすることも考えられる．

8 今後の課題

(1) 探索範囲限定アプローチのプロトタイプ実装と評価
(2) 探索範囲限定アプローチの評価に基づくプロトタイプ修正

9 まとめ

特許文書分析プロセスにおいて，特許文書分類（機械学習）前に文書フィルタにより探索範囲を発明のポイントとなる可能性が高い箇所に限定する探索範囲限定方法を提案した．本提案方法により，特許調査の効率化，機械学習処理効率向上，および特許調査における再現率向上が期待できる．

提案方法の試行評価から，探索範囲を限定するだけでなく，調査対象の請求項の分析および特許文書（先行技術文献）の分析の必要性を明らかにした．したがって，探索範囲限定方法に加え，これらの分析も行って特許文書の分類を行うことが望ましいと考える．

参考文献

[1] 安藤 俊幸 他，機械学習を用いた効率的な特許調査，第 14 回情報プロフェッショナルシンポジウム予稿集，No. A31，情報科学技術協会，Nov. 2017, pp. 1-6.
[2] 工業所有権協力センター，特許検索競技大会 過去問 2016, 2017.
[3] 工業所有権情報・研修館，特許分類の概要とそれらを用いた先行技術調査～IPC、FI、F ターム編～（平成 30 年度），2018, https://www.inpit.go.jp/jinzai/kensyu/kyozai/outlink00057.html.
[4] J. A. Lau, et al., An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, Proc. of the 1st Workshop on Representation Learning for NLP, ACL, Aug. 2016, pp. 78-86.
[5] M. Lupu, et al. (Eds.), Current Challenges in Patent Information Retrieval, 2nd ed., Springer, 2017.
[6] 鈴木 祥子，機械による特許分析の課題とアプローチ，情報の科学と技術，Vol. 67, No. 7, 2017, pp. 355-359.
[7] 特許庁，特許・実用新案審査基準 事例集，特許・実用新案審査ハンドブック 附属書A, Apr. 2019, https:// www.jpo.go.jp/system/laws/rule/guideline/patent/handbook_shinsa/index.html.