1J3-2

特許文献を利用した技術課題の抽象化方法の検討

Abstraction Method of Technical Problems Using Patent Documents

谷中 瞳[*1]
Hitomi Yanaka

大澤 幸生[*1]
Yukio Ohsawa

[*1] 東京大学大学院 工学系研究科 システム創成学専攻
Department of System Innovation, Faculty of Engineering, the University of Tokyo

In companies, patent analysis is important to understand the trends of technical problems and develop technology strategy and intellectual property strategy. Patent classification is a method of supporting patent analysis and the method of automatic patent classification is required. In this study, we propose a method for extracting information on the technical problem from patent documents by categorizing patent documents with the use of hierarchical clustering due to the similarity of the patent documents of "problem to be solved by the invention" and "means to solve the problem" and summarizing the technical problem of each patent document. We also propose the visualization of common technical problem and summary of the technical problem of each patent document.

1. はじめに

企業において、特許分析によって技術課題の動向を把握し、技術開発戦略や知財戦略を立案することは重要である。特許文献には専門的な技術用語が多く含まれているため、漏れやノイズのない分析方法が求められているが、特許分析を支援する方法として特許分類がある。技術内容に応じて特許文献に分類コードを付与する分類体系として、IPC（国際特許分類）という一般/具体関係の階層をもった分類体系や、国や地域による技術分野ごとの発明数を考慮するために考案された、FI、ECLA、USPC といった各国独自の分類体系がある。また、機能だけではなく目的、利用分野、材料といった多観点の分類体系としてFタームがある。これらの分類コードは専門家によって人手で付与されており、意味内容に基づいて自動で分類する技術が求められている。

また、技術課題と解決手段を軸として特許出願動向を可視化したものとしてパテントマップがあるが、パテントマップは技術動向を俯瞰的に把握する方法であり、マップから各クラスタに具体的にどの特許文献が属しているのかを一目で特定することができない。また、パテントマップも専門家によって人手で軸を指定し作成しているのが現状である。

そこで本研究では、企業の技術開発戦略の創出を支援する方法として、特許明細書の「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」内の文書に基づいてクラスタリングを行うことにより、クラスタ内の特許文献に共通する技術課題を抽出することを目的とする。さらに、クラスタに共通する技術課題と、各特許文献の技術課題の内容が一目で把握できるよう可視化することを目的とする。

連絡先：

谷中瞳, 東京大学大学院工学系研究科システム創成学専攻, h2.yanaka@gmail.com
大澤幸生, 東京大学大学院工学系研究科システム創成学専攻, ohsawa@sys.t.u-tokyo.ac.jp
This research is supported by JST, CREST. また、本研究を支援してくださっているトッパン・フォームズ株式会社の皆様に感謝申し上げます。

2. 関連研究

2.1 特許文献からの技術課題抽出

文書を自動分類してトピックを抽出する方法としては、新聞記事に含まれる単語を特徴量として階層的クラスタリングを行うことによってトピックを抽出する方法[橋本 08]が提案されている。しかし、先行研究はクラスタ数を人手で決定しているため、クラスタ数によって抽出されるトピックも変動してしまう問題がある。

また、特許文献から技術課題を抽出する方法としては、特許明細書内の項目「発明の効果」に含まれる文書から「～ができる」という手がかり表現をもとに新たな手がかり表現、手がかり表現から技術課題情報を示す文書として共通に頻出する表現、と順々に表現を獲得し、パテントマップの自動作成に必要な技術課題情報を抽出する方法[酒井 09]が提案されている。

本研究では、Upper Tail 法[志津 11][Mojena 77]により適切なクラスタ数を決定することで、抽出されるトピックの変動を解消する。また、各特許文献が解決しようとする技術課題の要点を抽出することを目的とするため、技術課題が端的に記載された項目「発明が解決しようとする課題」の内の文書を技術課題の抽出対象とする。

2.2 類推を用いたアイデア創出方法

過去の自社・他社の特許を分析し技術開発戦略を立案する上で、類推を用いたアイデア創出は有用である。類推とは、ベース（既知の状況）に関する知識や経験をターゲット（解決しようとする課題が存在している状況）に当てはめることによって、ターゲットにおける課題を理解し、解決するプロセスである。構造写像理論によると、類似性にはベースとターゲットに含まれる要素によって特徴づけられる表層的類似性と、ベースとターゲットに含まれる要素間における一次または高次の関係によって特徴づけられる構造的類似性がある[Falkenheiner 89]。表層的類似性に基づいてベースと共通する要素を検索し推論を行い、構造的類似性に基づいてベースとターゲットに含まれる要素間の関係を比較し、推論の妥当性を評価するモデルとして、MAC/FAC モデルが提唱されている [Forbus 94]。

そこで、技術課題の要素による表層的類似性に加えて、技術課題文の構造に基づく構造的類似性に基づいて技術課題を表現することによって、より推論を支援できる可能性がある。

また、類推を用いたデータ利活用に対する人間の創造性と価値発見のための支援手法として、データジャケット（以下、DJ）がある[Ohsawa 13]。DJ とは、データに含まれる変数やデータの形式などのデータの内容を説明するメタデータの一種である。データの中身が非公開でも、データの概要である DJ を公開することで、どこにどのようなデータが存在するのか理解でき、データから課題を検討できる。DJ を組み合わせたアイデアを出しやすくするための方法の 1 つとして、DJ の「動詞化」が提案されている[Miura 16]。DJ の「動詞化」とは、注目する DJ 内の変数を名詞格とする動詞を当てはめることにより、DJ の理解度を高める方法である。例として、「ETC の明細履歴情報」という DJ には、「車載シリアル、ゲート通過時刻、料金」という変数が含まれている。これらの変数と「特定する」という動詞を組み合わせると、「車の 1 日の ETC にかかった料金を特定する」という文書が作成できる。

DJ の「動詞化」のように、文章を述語論理形式で記述することによって、推論を促進しうる可能性がある。文章から述語と項の関係を捉えるためには、述語項構造解析が必要となる。述語項構造の体系化に関する研究として、格フレーム辞書[河原 05]がある。格フレーム辞書は用言とそれに関係する名詞を用言の用法ごとに整理したものであり、 これを活用することによってある特定の名詞に対応する動詞の候補を、また逆に、動詞に対応する名詞の候補を検索することができる。

本研究では、「発明が解決しようとする課題」の内の文書のうち、「本発明」を主語とした動詞と対応する目的格を技術課題として抽出することにより、各特許文献の技術課題に対する働きかけ方を示し、各特許文献の理解を支援する形で表現することを目指す。

3. 階層的クラスタリングによる共通課題の抽出

本研究では 2013 年から 2015 年に発行された公開公報のうち、「発明が解決しようとする課題」に「調味料」が含まれる特許文献 348 件を対象とし、技術課題を抽出する。提案手法は下記の 4 つの手順から成る。図 1 に提案手法の全体像を示し、各手順の詳細を以下に記述する。

Step.1 特徴量の導出とクラスタリング
Step.2 クラスタに属する特許文献の技術課題の抽出
Step.3 クラスタに属する特許文献に共通する技術課題（以下、共通課題と定義する）の抽出
Step.4 技術課題の可視化

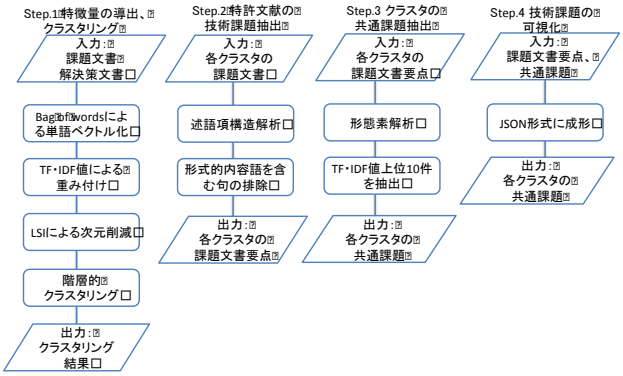

図 1. 提案手法の全体像

3.1 特徴量の導出とクラスタリング

本研究で「技術課題が類似する」とは、課題が類似する場合、解決策が類似する場合の両方を示すので、特許明細書のうち「発明が解決しようとする課題」と「課題を解決するための手段」の文書（以下、課題文書、解決策文書とする）から特徴量を導出する。

まず、課題文書・解決策文書中の名詞、動詞、形容詞の単語の出現頻度から bag of words を利用して特徴ベクトルを導出する。このとき、特許文献において一般的に使用される単語を除くため、辞書作成時に出現頻度 30%以上の単語はフイルタをかける。また、各特許文献に含まれる単語の重み付けを行うため、文書における単語の出現頻度(TF)と全文書における単語の出現頻度の対数(IDF)を掛け合わせた TF・IDF 値を各特許文献の特徴量とする。次に、高次元ベクトルをそのまま特徴量としてクラスタリングを行うと球面クラスタリングとなり分類の精度が落ちるため，文書ベクトルの次元を圧縮して扱う必要がある。今回は LSI（Latent Semantic Indexing）を用いて、各特許文献の特徴量を予備実験により決定した 200 次元に次元削減し、正規化を行う。

次に、各特許文献の教師なしクラスタリングの方法として、Ward 法による階層的クラスタリングを行う。階層的クラスタリングを採用することによって、クラスタ間の関連性を可視化に反映することができる。なお、特許文献間の距離はユークリッド距離を採用する。

次に、クラスタ数を検討する。階層的クラスタリングのクラスタ数自動決定法としては、Upper Tail 法[Mojena 77]が提案されている。Upper Tail 法は、n 件の文献を n-j(j=0,1,…,n-1)個のクラスタに分割するためのクラスタ間の距離（異なるクラスタに属する 2 文献間の最少距離）α_j について、式(1)の停止規則に基づいてクラスタ数を決定する方法である。式(1)を満たすまで j を増加させ、停止した j が最適なクラスタ数となる。α_{ave} は α_j の分布の平均、s_α は不偏分散の平方根、k は 1 クラスタあたりの文献数から決定される定数であり、本研究では 1 クラスタあたりの文献数を 10～50 件と想定していることを考慮して、k=2 とする。

$$\alpha_{j+1} > \alpha_{ave} + ks_\alpha \quad (1)$$

また、クラスタ分析におけるクラスタ数自動決定法の比較を行った研究[志津 11]によると、多変量正規分布に従う 2 点間の距離の分布はカイ 2 乗分布に従うことを考慮して、式(2)のように α をカイ 2 乗分布に従うよう正規化してから Upper Tail 法を用いるほうが、精度が良いとされている。

$$\alpha' = \Phi^{-1}(F_p(\alpha/s_\alpha \cdot p) \quad (2)$$

ここで、Φ^{-1} は正規分布の逆関数であり、Fp は自由度 p のカイ二乗分布の分布関数である。本研究ではこの手法を用いて適切なクラスタ数を求める。

3.2 特許文献の技術課題抽出

各クラスタに共通する技術課題を抽出するためには、クラスタに属する各文献が解決しようとする技術課題の要点を抽出する必要がある。課題文書の記述方法は出願者によって様々であるが、一般的には「本発明は…する」という表現を用いて各特許文献が解決しようとする技術課題の要点を記述している。そこで、「本発明は…する」という表現を手がかり表現として、述語項構造解析ツールを用いて「本発明」を主語とする動詞と目的格を抽出する。課題文書は短い文書から長い文書まで多岐に渡るため、格フレーム辞書[河原 05]に基づくルールベースの述語項構造解析ツール KNP を採用する。

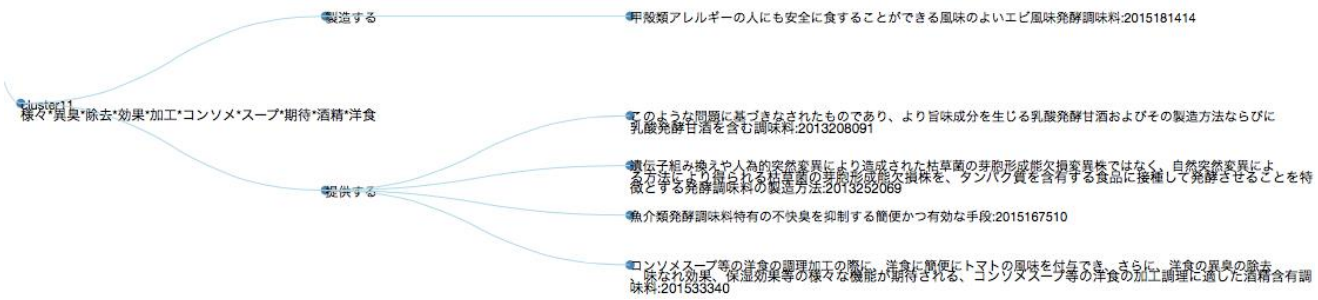

図 2. 「調味料」に関する技術課題と関連する特許の技術課題要約の可視化(一部)

なお、「本発明」を主語とした文章の多くは「ものである」「こととする」といった意味を持たない語の組み合わせを同じ句に持つ。構文片に関する研究[瀧川 11]ではこの「もの」「こと」といった語を「形式的内容語」と定義しており、これらの語を含む句を取り除く処理を行う必要がある。そこで、抽出した技術課題から人手で収集した形式的内容語「もの」「こと」「事」「問題」「課題」「目的」を名詞格に持つ動詞とその名詞格を取り除く。

以上より、各特許文献の課題文書のうち、「本発明」から始まる文章を技術課題の要点を含む文章と仮定し、文章から動詞と目的格を抽出し、形式的内容語を含む句を取り除いたものを技術課題の要点とする。

3.3 クラスタの共通課題抽出

次に、クラスタに属する各文献の技術課題の要点から、クラスタの共通課題を抽出する。クラスタに属する各文献の技術課題の要点で使用されている名詞の出現頻度(TF)と、全文献の技術課題の要点で使用されている名詞の出現頻度の対数(IDF)を掛け合わせた TF・IDF 値を算出する。本研究では、TF・IDF 値が大きい単語を共通課題とするため、0.1 以上の単語を共通課題の候補とし、このうち上位 10 件までをクラスタの共通課題と定義する。

3.4 技術課題の可視化

最後に、クラスタ番号と各クラスタの共通課題、各クラスタの特許文献の技術課題で使用されている動詞、その動詞に対応する名詞格と公報番号の順にネスト構造を作成し、プログラミング言語間で交換可能なデータ形式である JSON 形式で出力する。ツリー構造の可視化にはデータ可動型の可視化ライブラリである D3.js を用いる。

4. 結果と考察

4.1 実験結果

3.1 の方法で最適なクラスタ数を計算した結果、クラスタ数は 14 個となった。また、可視化した図の一部を図 2 に示す。

4.2 評価

技術課題に基づいてクラスタリングできていることの妥当性について、テーマコードの適合率によって評価した。テーマコードとは特許文献を技術範囲ごとに区分するための記号であり、各文献に 1～4 個のテーマコードが付与されている。そこで、式(3)に基づいて各クラスタのテーマコードの適合率を算出した。T はクラスタ i 内の文献で最も多く付与されていたテーマコード(以下、代表テーマコード)、$N_{i,T}$ はクラスタ i 内の文献のうち T が付与されている文献数、$N_{i,All}$ はクラスタ i に属する全文献数である。

$$p_{i,T} = N_{i,T} / N_{i,all} \quad (3)$$

各文献の技術課題の要点の妥当性は、抽出できた要点のうち無作為に選んだ 100 件と、食品に関する特許の有識者 1 名と学生1名に依頼し、同じ特許文献の課題文書に対して人手によって技術課題の要点を抜き出した 100 件とを比較し、同じ文を取得している割合を合致率として評価した。なお、依頼者に「本発明」という手がかり表現については伝えていない。

最後に、クラスタから抽出された共通課題の妥当性を検討するため、食品関連の研究開発業務および出願・権利化業務に 10 年以上従事している専門家 3 名にヒアリングを行い、本研究で作成した図を評価してもらった。評価の観点は下記の 3 点であり、それぞれの観点による評価とその理由をヒアリングした。

①クラスタリングの結果は技術課題に基づいて適切にクラスタリングされているか

②各特許文献の技術課題の要約は把握できる内容であり、先行技術調査において有用性があるか

③各クラスタから抽出された共通課題は把握できる内容であり、先行技術調査において有用性があるか

4.3 考察

課題文書を特徴量としたクラスタリングについて考察する。テーマコードの適合率によって各クラスタの妥当性を評価した結果を表 1 に示す。

表 1. 各クラスタのテーマコードの適合率

クラスタ番号	文献数 / 件	代表テーマコード	適合率 / %
0	75	4B036	13.3
1	27	4B047	96.3
2	22	4B047	77.3
3	12	3E084	91.7
4	12	4B047	41.7
5	18	4B047	83.3
6	74	4B047	74.3
7	22	4B047	72.7
8	14	3L089	35.7
9	18	4B047	83.3
10	22	3E084	18.2
11	10	4B047	80.0
12	8	3E035	75.0
13	14	3L020	28.6

結果を見ると、14 件中 9 件が適合率 50%以上、5 件が適合率 50%未満であったため、適合率が低いクラスタについて考察する。クラスタ番号 4,8,10,13 については、1～3 番目に多く付与されたテーマコードについて適合率の合計を算出すると 83.3%,85.7%,54.6%,57.1%となり、3 番目に多く付与されたテーマコードまで考慮すると関連する技術課題に基づいてクラスタリングされていると考えられる。クラスタ番号 0 については、3 番目に多く付与されたテーマコードまで考慮しても 33.3%と適合率は低かった。

4.2 の①の観点に基づくヒアリングにおいても、クラスタ番号 0 については技術課題が共通していないという評価であり、クラスタ番号 0 に含まれる文献を見ると、「このような従来の技術が有する問題点に鑑みなされたものであり、」といった、技術課題を表現しない文節を含む課題文書が多く見られた。これは文書中の単語から特徴量を設計していることが原因であり、より文書の意味内容を反映した特徴量の設計が必要であることを示唆している。

次に、各特許文献から抽出された技術課題について考察する。まず、課題文書の要点が必ずしも「本発明は…する」という形式で記述されていない場合があるため、特許文献の技術課題を抽出できた例は 188 件/348 件(54.0%)に留まった。また、「本発明は…する」という形式で記述されている場合でも、下記に抽出に成功した事例と失敗した事例を示すように、意味を持たない語句が抽出されてしまい失敗する事例があった。

技術課題の要点の抽出に成功した事例：サンドイッチの包装を片手で容易に開封して手早く食べ始めることができるサンドイッチ用包装袋を/提供する

技術課題の要点の抽出に失敗した事例：以上の問題点を/解決するためになされた

これは「問題点」という単語を形式的内容語に含めていなかったことが原因であり、形式的内容語の範囲を広げることによって改善できると考えられる。

また、抽出できた要点と人手で作成した要点とを比較した結果、合致率は 94.0%であり、「本発明」は手がかり表現として妥当であることが示唆された。

4.2の②の観点に基づくヒアリングにおいても、抽出された技術課題の要点は上記の失敗例を除いては十分に理解できる内容であり、どんな技術分野の文献か一目でわかるため有用であると判断された。また、3 名中 2 名の評価者から課題だけではなく解決方法も抽出するとなおよいという意見があり、これは今後の課題とする。

最後に、クラスタの共通課題について 4.2 の③の観点に基づくヒアリングの結果から考察する。ポジティブな評価としては、共通課題からどのようなタイプの食品に関する内容かが一目でわかり、初めて取り組む技術テーマを扱うときに見やすい、クラスタ番号によっては加熱なのか分解なのかといった、素材だけではなく加工方法についての単語も抽出できているという評価があった。また、共通課題の単語の数は適切であると評価された。ネガティブな評価としては、前述のクラスタ番号 0 など、クラスタ番号によっては適切な共通課題が抽出されていない場合があり、ある程度検索キーワードがはっきりしている製品開発段階での調査では共通課題の精度が求められるため、適切ではないという評価があった。この結果から、共通課題を抽出することの有用性が示唆された一方で、より精度の高い抽出方法を検討する必要があると考えられる。このとき、抽出される共通課題はその前段階のクラスタリング、技術課題の抽出によって左右されるため、前述の特徴量の設計方法を見直すことにより改善されると考えられる。

5. まとめ

本研究では、特許明細書の課題文書と解決策文書中の単語の類似度に基づきクラスタリングを行い、技術課題を抽出する方法を検討した。ヒアリングの中で、自分が調べたい課題について一覧で内容がつかめるのでとても良いという評価もあり、提案手法による技術課題の可視化は、ユーザーが技術課題に関連する特許を検索する時の支援として有効であると考えられる。

今後の課題としては、より創造的な技術課題戦略の提案を支援するために、構造的類似性を考慮したクラスタリングを行うことが挙げられる。今回は課題文書、解決策文書中の単語を特徴量としてクラスタリングを行ったため、表層的類似性に基づくクラスタリングとなっている。文書中の単語そのものではなく文書の意味内容に基づいて述語論理形式で記述し、得られた述語を特徴量とすることによって上記の実現に取り組む予定である。

参考文献

[橋本 08] 橋本 泰一, 村上 浩司, 乾 孝司, 内海 和夫, 石川 正道, 文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集, Vol.5, p.216-226, 2008.

[酒井 09]酒井 浩之, 野中 尋史, 増山 繁, 特許明細書からの技術課題情報の抽出, 人工知能学会論文誌, Vol.24, No.6 p.531-540, 2009.

[志津 11] 志津綾香, 松田眞一, クラスター分析におけるクラスター数自動決定法の比較, アカデミア, 情報理工学編, Vol.11, p.17-34, 2011.

[Mojena 77] Mojena, R., Hierarchical grouping methods and stopping rules: an evaluation, The Computer Journal, Vol.20, p.359-363, 1977.

[Falkenheiner 89] Falkenheiner, Forbus, Dedre Gentner, The Structure Mapping Engine: Algorithm and Examples, Artificial Intelligence 41, p.1-63, 1989.

[Forbus 94] Forbus, Gentner, and Law, MAC/FAC: A model of similarity-based retrieval, Cognitive Science, 19, p.141- 205, 1994.

[Miura 16] Miura, D., Ohsawa, Y., K. Furuta, Predicate-based knowledge elicitation for action, 人工知能学会全国大会, 2016.6.

[Ohsawa 13] Ohsawa, Y., Kido, H., Hayashi, T., Liu, C., Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013, Procedia Computer Science 22, p.709-716, 2013.

[河原 05] 河原 大輔, 黒橋 禎夫, 格フレーム辞書の漸次的自動構築 自然言語処理, Vol.12, No.2, p.109-131, 2005.

[瀧川 11] 瀧川 和樹, 山本 和英, 構文片の改良と評判分析への適用, 言語処理学会第 17 回年次大会発表論文集, p.111-114, 2011.