The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

特許俯瞰解析のためのクラスタラベリング手法

Clueter Labeling for Patent Panoramic Analysis

尾崎 花奈 *1　　十河 泰弘 *1
Kana Ozaki　　Yasuhiro Sogawa

*1日立製作所 研究開発グループ
Research & Development Group, Hitachi, Ltd.

Patent data is important for companies to grasp the technology trends and the status of competitors. Patent panoramic analysis supports companies to position themselves in the market and grasp the trends of competitors, because it enables users to overview patents in a specified technical field by clustering and visualizing them. In each cluster, the users can capture the technical features of the patents by showing them cluster labels (representative keywords in the cluster). However, cluster labels often overlap among the clusters when we try to assign labels which consists of high frequency words in each cluster. This is because words that appear frequently in one cluster also appear frequently in other clusters when we cluster patents in a specified technical field. To tackle this challenge, we propose to use phrases as cluster labels and extract optimal combinations of the phrases which are discriminative among the clusters. In the experiment, we evaluate the effectiveness of our proposed method by using patent datasets.

1. はじめに

特許情報は，企業が知財戦略策定を行ううえで重要な情報である．特に近年，経済のグローバル化や企業の競争激化に伴い，特許情報を用いて自社保有特許群の位置づけや競合他社の技術開発動向を分析し，分析結果を知財戦略だけでなく事業戦略や研究開発戦略に活用することが必要になっている．このような知的情報の分析結果を自社の経営戦略に役立てる取り組みは IP ランドスケープと呼ばれ [杉光 19]，日本国内において大手企業を中心に取り組みが広まっている．IP ランドスケープを支える代表的な技術として，特許の俯瞰解析が挙げられる．俯瞰解析では，着目する技術分野の特許集合における特許俯瞰マップを作成することで，技術トレンド，自社の位置づけ，他社が競合しうる技術領域を効率的に把握・考察することが可能である．

特許俯瞰マップは，特許公報 1 件 1 件を 1 データ点として捉え，可視化によってどのようなまとまりが存在しているかという情報を直感的に把握することを促す [岩本 15]．マップ作成は，各特許公報をベクトル化し，クラスタリングを行うことで技術分野の細分化を行い，各クラスタを代表するキーワードを表示するという流れになっている．キーワード表示のように各クラスタの特徴を表すものはクラスタラベルと呼ばれ，ラベルを生成しユーザに提示することはクラスタラベリングと呼ばれる．図 1 にクラスタラベルを伴う特許俯瞰マップのイメージを示す．クラスタ内における特許公報の直感的な理解を促すためには，他クラスタの特許公報と差別化され，かつ，クラスタ内の特許公報全体を網羅する，可読性の高いクラスタラベルの表示が求められる．この点において，単純にクラスタ内での出現頻度を基に選択した単語をラベルとする場合では，以下 2 つの課題が存在する．

1. ある着目する技術分野に対して特許マップを作成するため，各クラスタで高頻度で出現する単語は重複する傾向にあり，クラスタラベルがクラスタ間で重複してしまう．
2. クラスタ内で高頻度で出現する単語で構成されたクラスタラベルは，代表的な 1 部の技術のみを表しており，クラスタ全体の内容を網羅していない．

これらの課題を解決するため，本稿では，他クラスタとの重複を防ぎながらクラスタ内の内容網羅性の高いラベルを生成するクラスタラベリング法を提案する．

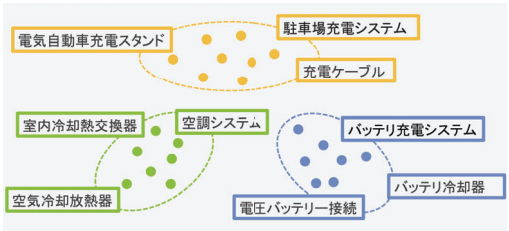

図 1: クラスタラベルを伴う特許俯瞰マップのイメージ

2. 関連研究

クラスタラベリングと関連して，テキストに対してトピックモデルによりトピック解析を行った結果に対してラベル付けを行うトピックラベリングの研究が盛んに行われている．トピック解析では，各トピックに対して単語の出現分布を作成するが，文書をクラスタリングした場合にも，各クラスタにおいて単語の出現分布は容易に求めることができるため，トピックラベリングはクラスタラベリングとして応用することが可能である．

Mei ら [Mei 07] は生成するラベルとトピック間の KL ダイバージェンスを最小化しながら自己相互情報量を最大化する問題を解くことで，トピックを代表するラベルを生成する方法を提案している．Baseve ら [Cano Basave 14] は，Twitter において抽出したトピックに対して，TF-IDF(Term Frequency-Inverse Document Frequency) や Text Rank などの既存の要約アルゴリズムを用いてラベルとなる単語の羅列を選択する方法を提案している．Wan ら [Wan 16] は，トピックごとにラベルとなる複数の文を，トピックと関連性があり，トピック

連絡先: 尾崎花奈，日立製作所，〒 185-8601 東京都国分寺市東恋ヶ窪一丁目 280 番地，kana.ozaki.dj@hitachi.com

内の内容を網羅し，他トピックのラベルと差別化されるように選択する文抽出型要約によるラベリングを，劣モジュラ最適化アルゴリズムを用いることで実現している．これらの手法は，外部知識を用いず，クラスタリング対象の文書のみを用いてラベルを生成することができる．

一方で，外部知識を用いて，ラベルの情報量を増やす方法も提案されている．Lau ら [Lau 11] は，英語 Wikipedia を用いて，各トピックの上位単語を含む記事のタイトルからラベル候補を生成する方法を提案している．Hulpus ら [Hulpus 13] は，DBpedia の構造化データを利用することで，トピックに関連するトピックグラフを抽出し，抽出したグラフに対してラベリングを行う方法を提案している．

3. 提案手法

本稿では，Wan らの文抽出型要約によるラベリング手法をもとにして特許公報から抽出した複合語を対象にラベリングを行うことで，他クラスタのラベルと差別化を図りながらクラスタ内を網羅するラベルの生成を行う．特許公報においては 1 文が長い傾向にあるため，Wan らの手法における文形式のラベルは冗長であり，かつ，文中での情報量が多いためにクラスタの内容を限定しすぎてしまう．そこで本研究では，複合語をラベルの対象とすることでラベルの冗長性を防ぎながら，単語レベルよりも可読性の高いラベル生成を実現する．

3.1 複合語抽出

クラスタごとに，特許公報から専門用語抽出ツール termextract[中川] を用いて複合語を抽出する．termextract は，名詞 (単名詞と複合名詞) を対象として，頻度と左右単語との連接情報から専門用語としての重要度を計算する．本稿では，termextract で抽出された専門用語のうち，1 つの単語から成るものは削除し，かつ，重要度上位 100 複合語を候補ラベルとしてラベリングの対象とする．

3.2 ラベル最適化

Wan らの手法に倣い，3.1 で抽出された候補ラベル集合 V において，クラスタ内の特許公報集合と関連し，クラスタ内の全体の内容を網羅し，他のクラスタラベルとは異なるような複合語集合 $\widetilde{E}$ を，スコア関数 $f(E)$ を複合語の個数 ($|E|$ とする) の制約の中で最大化することで最適化していく．

$$\widetilde{E} = \arg\max_{E \subseteq V} f(E) \quad s.t.\ |E| < L \tag{1}$$

L は複合語個数の制約を表す．スコア関数 $f(E)$ は，以下の 3 つの項から構成される．

$$f(E) = REL(E) + COV(E) + DIS(E) \tag{2}$$

これら 3 つの項，$REL(E)$，$COV(E)$，$DIS(E)$ は，それぞれ前述の，関連性，網羅性，差別化を表すスコアであり，それぞれのスコア関数について以下で説明する．

関連性スコア

$$REL(E) = \sum_{t' \in V} \min\{\sum_{t \in E} sim(t', t), \alpha \sum_{t \in V} sim(t', t)\} \tag{3}$$

$sim(t', t)$ は複合語 t' と複合語 t のコサイン類似度であり，$\sum_{t \in E} sim(t', t)$ は選択した複合語の集合がどれだけ複合語 t' に類似しているかを表している．$0 \leq \alpha \leq 1$ は閾値を調整するパラメータである．$\sum_{s \in V} sim(t', t)$ は，$\sum_{s \in E} sim(t', t)$ が到達可能な最大値を表す．$\min\{\sum_{t \in E} sim(t', t), \alpha \sum_{t \in V} sim(t', t)\} = \sum_{t \in E} sim(t', t)$ となったとき，t' に関して E は飽和状態にあることを示すため，t' と類似する複合語はこれ以上 $REL(E)$ を改善させることはなく，t' と類似しない複合語のみ $REL(E)$ を改善することができる [Lin 11]．このように $REL(E)$ はなるべくクラスタ内の文書集合全体と類似するように複合語を選択していく．

網羅性スコア

関連性スコアだけでも，ある程度文書集合全体を網羅するようなラベルを生成可能だが，以下の式のように選択される複合語集合が多様性を持つことに関して報酬を与えることで，網羅性を向上させることが可能である．

$$COV(E) = \beta * \sum_{w \in TW} \left\{ p_\theta(w) * \sqrt{\sum_{t \in E} tf(w, t)} \right\} \tag{4}$$

$\beta \geq 0$ は結合係数であり，TW はクラスタ内の頻度や TF-IDF の重みが上位の単語集合である．$p_\theta(w)$ はクラスタ θ における単語出現確率や TF-IDF の重みを表しており，$tf(w, t)$ は，複合語 t 中の単語 w の頻度を表す．$f(x) = \sqrt{x}$ の性質から，既に選択された複合語 t_1 と類似した複合語 t_2 よりも類似しない複合語 t_3 を選択した方が網羅性スコアが高くなる傾向にある ($COV(t_1, t_2) < COV(t_1, t_3)$)[Lin 11]．このように，上位単語 TW の中からなるべく多くの単語を含む複合語集合が選択されることに対して報酬を与える．

差別化スコア

該当クラスタ θ におけるラベル（複合語集合）E と，その他のクラスタ集合 $\{\theta'\}$ との差異を表現したスコア関数が以下である．

$$DIS(E) = -\gamma \sum_{\theta'} \sum_{t \in E} \sum_{w \in TW} p_{\theta'}(w) * tf(w, s) \tag{5}$$

$\gamma \geq 0$ は結合係数である．負の符号にすることで，E がその他のクラスタの上位出現単語を含むことに対してペナルティを与えている．これによって，異なるクラスタに対して異なるラベルが選択されるように調整される．

4. 実験評価

4.1 実験設定

特許母集団データとして，特許庁が公開する特許公報データより，全文中に「電気自動車」「充電」を含む 2021 年 1 月から 3 月に公開公報に掲載された日本語特許データ 1207 件数を使用した．

クラスタリングを行うために，TF-IDF 法により各特許公報中に含まれる名詞の重みを計算し，ベクトル化した．この際,「装置」「システム」といった全特許公報共通の頻出単語 42 単語をストップワードとしてあらかじめ除外したうえで，特許公報母集団全体で DF (Document Frequency) が上位 90% である単語と，下位 5% である単語は除外した．日本語の単語分かち書きには MeCab を用いた．クラスタリング法には，K-means 法を採用し，文書間の類似度はコサイン類似度により計算し，クラスタ数は 5 と設定した．

クラスタラベリングは，上記のクラスタリングで得られた結果に対して行い，複合語の抽出対象とする文は,「要約」部分に

限定した．網羅性スコアと差別化スコアの計算に用いる $p_\theta(w)$ は，クラスタリング時に計算した各単語の TF-IDF 値とした．パラメータ設定は，Wan らの手法に従い表 1 のようにした．

表 1: パラメータ設定

パラメータ	概要	値
L	ラベルとする複合語の個数	5
TW	ラベル最適化に用いる単語数	50
α	関連性スコアのパラメータ	0.05
β	網羅性スコアのパラメータ	250
γ	差別化スコアのパラメータ	300

比較手法として，以下のラベリング手法を設定し，提案手法に合わせて各クラスタ 5 つのラベルを出力した．

TF-IDF 上位単語ラベル
各単語の TF-IDF 値をクラスタごとに集計し，上位 5 単語をラベルとする．

TF-ICF 上位単語ラベル
1 クラスタを 1 文書とみなし，IDF の代わりに ICF (Inverse Cluster Frequency) を計算し，TF-ICF 値が上位の 5 単語をラベルとする．多くのクラスタで高い頻度で出現する単語の TF-ICF 値は低くなるため，ラベルとして選ばれにくくなる．

TF-IDF 上位複合語ラベル
提案手法と同じ方法で抽出した各クラスタ 100 複合語を候補ラベルとして，複合語内の各単語 TF-IDF 値の平均をその複合語の重みとして、上位 5 複合語をラベルとする．

評価においては，ラベルのクラスタ間重複度とクラスタ内網羅性を測定した．重複度は，2 クラスタ間のラベル同士の Simpson 係数を全組合せで平均したものを評価値とした．ある 2 つの異なる集合 X と Y の Simpson 係数 $Simpson(X, Y)$ は以下の式で求められ，2 つの集合が完全に一致しているときに 1，共通部分を持たないときに 0 となる．

$$Simpson(X, Y) = \frac{|X \bigcap Y|}{\min(|X|, |Y|)} \tag{6}$$

クラスタ間でラベルの重複が少ないほど良いため，重複度は小さい方が良い評価値となる．網羅性は，クラスタ内上位 50 単語のうち各クラスタで生成されたラベルがそれらの単語を含む割合を全クラスタで平均したものを評価値とした．網羅性は大きい方が良い評価値となる．なお，単語単位のラベルにおいては 5 単語から成るラベルであるので，全てのクラスタにおいて網羅性評価値は $5/50 = 0.1$ となる．

4.2 結果

各手法で得られたラベルのクラスタ間重複度とクラスタ内網羅性を表 2 に示す．重複度，網羅性ともに提案手法が最も良い結果となっている．重複度については，単語単位のラベルでクラスタ間の重複が見られるのに対して，提案手法を含む複合語単位のラベルでは重複が 1 つも無い．TF-ICF 上位単語ラベルは，TF-IDF 上位単語ラベルと比較して重複が減っていはいるが，0 ではない．また，網羅性については，TF-IDF 上位複合語ラベルが，単語単位のラベルと比較してほとんど向上していないのに対して，提案手法では大幅に向上している．

表 2: クラスタ間重複度とクラスタ内網羅性

手法	重複度	網羅性
TF-IDF 上位単語ラベル	0.08	0.1
TF-ICF 上位単語ラベル	0.06	0.1
TF-IDF 上位複合語ラベル	**0.0**	0.108
提案手法	**0.0**	**0.196**

4.3 考察

各手法において出力されたラベルを表 3 に示す．TF-IDF 上位単語ラベルは,「バッテリ」「電池」「電力」「制御」が複数クラスタに出現している．特にクラスタ 3 とクラスタ 5 は,「電力」「制御」が両クラスタに出現しているので，2 つのクラスタの内容の差を捉えることが難しい．TF-ICF 上位単語ラベルもおいても,「電池」「電力」「制御」が複数クラスタに出現している．TF-IDF 上位単語ラベルからラベルの内容は少し変化しているが，同じようにクラスタ 3 とクラスタ 5 の内容の差を捉えることができない．また，これら 2 つの単語ラベルは，単語単位のラベルであるために，何の技術を示しているのかが分かりづらい．例えば，クラスタ 4 に「部材」「方向」が含まれているが，何の部材なのか，何の方向なのかという情報がないため，これらの単語単体ではあまり意味を成さない．クラスタ 1 に出現する「交換」やクラスタ 5 に出現する「制御」も同様に，あまり意味を成さない．

一方で，TF-IDF 上位複合語ラベルは，特許公報に含まれる専門用語をラベルとしているため，単語ラベルと比較して情報量が大幅に増加している．単語ラベルにおいて，クラスタ 2 とクラスタ 4 では,「電池」という単語が共通して出現していたが，複合語ラベルを見ると，クラスタ 2 はリチウムに関する特許公報が集まっており，クラスタ 4 は電池に関する特許公報が集まっていることが分かる．また，単語ラベルにおいて，クラスタ 3 とクラスタ 5 では，単語ラベルでは区別が紛らわしかったが，クラスタ 3 は電圧に関する特許公報が集まっていて「制御」は電圧の制御の意味であり，クラスタ 5 は車両情報に関する特許公報が集まっていて「制御」は車両の制御の意味であるということが分かる．このように，TF-IDF 上位複合語ラベルにしたことで，ラベルの情報量が増え，各クラスタの特徴を捉えたクラスタ間の重複が無いラベルを生成できている．しかし，TF-IDF 値が高い単語を含むように複合語を選択したため，各クラスタのラベルは，複数の複合語に同一の単語を含んでいる．例えば，クラスタ 2 においては「リチウム」，クラスタ 3 においては「電圧」，クラスタ 4 においては「電池」が 5 つ全ての複合語に含まれている．このため，TF-IDF 複合語ラベルは，代表的な 1 部の技術のみを表しており，クラスタ全体の内容を網羅しているとは言えない．

これに対して，提案手法のラベルは，5 つの複合語が幅広い技術を網羅している．クラスタ 2 においては,「リチウム」を含む複合語だけでなく,「炭素原子」や「電極構造」といった複合語も含まれているため，電池の材料や電極構造に関する特許公報も含まれることが分かる．クラスタ 3 においては,「電圧」を含む複合語だけでなく,「直流電源」や「検出端子」といった複合語も含まれているため，電源装置や電流検知のシステムに関する特許公報も含まれることが分かる．クラスタ 4 においては,「電池」を含む複合語だけでなく,「端子支持部材」や「接続電極」といった複合語も含まれているため，電池周りの部品や電池の接続に関する特許公報も含まれていることが分かる．

表 3: 各手法で出力されたラベル

クラスタ	文書数	TF-IDF 上位単語ラベル	TF-ICF 上位単語ラベル	TF-IDF 上位複合語ラベル	提案手法
1	71	冷媒，バッテリ， 冷却，交換，配管	バッテリ，冷却， 冷媒，温度，交換	冷媒配管，バッテリ冷却モード， 冷却配管，バッテリ温度， バッテリ冷却システム	バッテリ冷却システム，冷媒温度センサ， バッテリユニット，車両搭載発熱機器， 空気循環
2	331	物質，リチウム， 電解，電池，粒子	物質，電池，リチウム， 電解，粒子	リチウム金属電池，リチウムイオン電池， 固体リチウム，金属リチウム， リチウム金属	リチウム含有アノード，リチウム金属電池， 金属複合水酸化物粒子，炭素原子， 電極構造
3	151	電圧，スイッチング， 回路，電力，制御	電圧，電力，制御， 回路，電流	電圧回路，制御電圧，電圧制御回路， 電圧変換回路，電圧変換制御	電圧回路，電流電圧変換回路， 制御電力，直流電源，検出端子
4	308	電池，電極，部材， セル，方向	電池，電極，方向， 接続，部材	電池セル，半導体電池，電池素子， 電池モジュール，電池パック	電池素子，端子支持部材，接続電極， 電池セル，回路基板
5	346	車両，情報，制御， 電力，バッテリ	車両，情報，制御， 電力，処理	車両情報，制御情報，給電車両， 車両制御ｅｃｕ，車両通知情報	電力システム，車両支持ユニット， 系統電力情報，自動運転車両，受電機器

このように，提案手法のラベルは，クラスタ内全体の内容を網羅したものになっている．また，提案手法では，3.2 における網羅性スコアや差別化スコアのパラメータを調整することで，クラスタ間の重複を避けることを重視したり，クラスタ内を幅広く網羅することを重視したりと，ラベルの出力をコントロールすることが可能である．

5. 終わりに

本稿では，特許俯瞰解析における特許公報集合のクラスタリング結果に対して，複合語によりラベリングを行う方法を提案し，クラスタ間でラベルの重複が無く，クラスタ内の網羅性が高いラベルの生成を試みた．日本語特許データを用いた実験によって，単語単位のラベルと比較して，クラスタ間で重複がないラベルを生成できていることを確認した．また，TF-IDF 値の平均が上位の複合語ラベルと比較して，クラスタ内の全体の内容を網羅したラベルを生成できていることを確認した．今後は，本稿のアプローチを拡張し，単語の分散表現を活用して単語の意味的関係性を考慮することで，より可読性の高いラベルの生成に取り組む予定である．

参考文献

[Cano Basave 14] Cano Basave, A. E., He, Y., and Xu, R.: Automatic Labelling of Topic Models Learned from Twitter by Summarisation, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 618–624, Association for Computational Linguistics (2014)

[Hulpus 13] Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D.: Unsupervised Graph-Based Topic Labelling using DBpedia, in *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pp. 465–474, ACM (2013)

[Lau 11] Lau, J. H., Grieser, K., Newman, D., and Baldwin, T.: Automatic Labelling of Topic Models, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1536–1545, Association for Computational Linguistics (2011)

[Lin 11] Lin, H. and Bilmes, J.: A Class of Submodular Functions for Document Summarization, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 510–520, Association for Computational Linguistics (2011)

[Mei 07] Mei, Q., Shen, X., and Zhai, C.: Automatic Labeling of Multinomial Topic Models, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 490–499 (2007)

[Wan 16] Wan, X. and Wang, T.: Automatic Labeling of Topic Models Using Text Summaries, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2297–2305, Association for Computational Linguistics (2016)

[岩本 15] 岩本圭介：特許情報テキスト可視化のためのマイニング手法, *Japio YEAR BOOK 2015*, pp. 276–279 (2015)

[杉光 19] 杉光 一成：IP ランドスケープ総論〜定義に関する一考察〜, 情報の科学と技術, Vol. 69, No. 7, pp. 282–291 (2019)

[中川] 中川裕志：専門用語（キーワード）自動抽出 Python モジュール termextract, `http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract/`：最終アクセス日：2022/2/17