ブロックチェーン
提案手法
実験

特許俯瞰解析のためのクラスタラベリング手法

海野　幸也 (Yukiya Unno)
u120006@st.pu-toyama.ac.jp

富山県立大学 情報システム工学科

December 5, 2023

はじめに

ブロックチェーン
提案手法
実験

背景

近年，経済のグローバル化や企業の競争激化に伴い，特許情報を用いて競合他社の技術開発動向を分析し，分析結果を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，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．また, 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．

ブロックチェーン
提案手法
実験

特許俯瞰マップ

特許俯瞰マップは，特許公報を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．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．

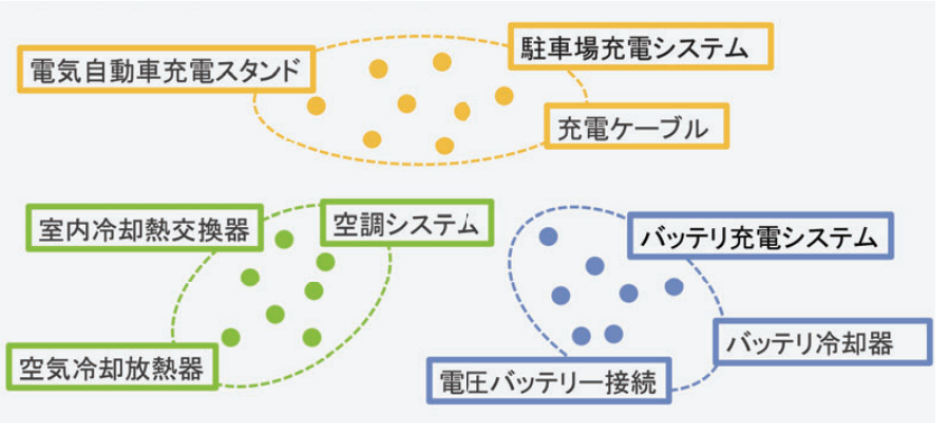

図 1: 特許俯瞰マップの例

提案手法

ブロックチェーン
提案手法
実験

提案手法の概要

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．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．

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 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．

提案手法

ブロックチェーン
提案手法
実験

ラベル最適化

複合語抽出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 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．

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

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．

$$f(E) = REL(E) + COV(E) + DIS(E)$$

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る．

提案手法

ブロックチェーン
提案手法
実験

関連性スコア

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

$sim(t', t)$ は複合語 t' と複合語 t のコサイン類似度であり，$\sum_{t' \in V} sim(t', t)$ 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．α は閾値を調整するパラメータである．$\sum_{t \in V} sim(t', t)$ は，$\sum_{t \in E} sim(t', t)$ が到達可能な最大値を表す．

提案手法

ブロックチェーン
提案手法
実験

網羅性スコア

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．

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

β 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．$p_\theta(\omega)$ 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，$tf(\omega, t)$ は，複合語 t 中の単語 ω の頻度を表す．平方根の性質から，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．

提案手法

ブロックチェーン
提案手法
実験

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.

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

γ 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.

実験

ブロックチェーン
提案手法
実験

実験設定

- 特許母集団データとして，特許公報データより，全文中に「電気自動車」「充電」を含む日本語特許データ 1207 件数を使用した．
- クラスタリングを行うために，TF-IDF 法により名詞の重みを計算し，ベクトル化した．
- 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．
- 羅性スコアと差別化スコアの計算に用いる $p_\theta(\omega)$ 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，図 2 のようにした．

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

図 2: パラメータ設定

実験

ブロックチェーン
提案手法
実験

比較手法として，以下のラベリング手法を設定した.

1. **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.
2. **TF-ICF 上位単語ラベル**
   クラスタを 1 文書とみなし，IDF の代わりに ICF を計算し，TF-ICF 値が上位の 5 単語をラベルとする.
3. **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.

ブロックチェーン
提案手法
実験

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図 3 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．

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**

図 3: クラスタ間重複度とクラスタ内網羅性

ブロックチェーン
提案手法
実験

考察

各手法において出力されたラベルを図 4 に示す.

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

図 4: 各手法で出力されたラベル

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．

まとめ

ブロックチェーン
提案手法
実験

まとめ

・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.
・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.