重要技術語を対象とした 特許技術の時系列トレンド分析手法 Patent-GLIPICA の開発

井畑 匠越 [1]　邊土名 朝飛 [2]　河野誠也[3]　原川 良介 [1]　岩橋 政宏 [1]　野中 尋史 [4]
[1] 長岡技術科学大学大学院工学研究科　[2] サイバーエージェント
[3] 理化学研究所 GRP　[4] 愛知工業大学
s225041@stn.nagaokaut.ac.jp　hentona_asashi@cyberagent.co.jp
seiya.kawano@riken.jp　{harakawa,iwahashi}@vos.nagaokaut.ac.jp
hnonaka@aitech.ac.jp

概要

本研究では，特許情報に基づく技術トレンド分析手法として，特許文書中の重要技術要素を抽出し，それらの時系列連動性を分析する手法「Patent-GLIPCA」を提案する．「Patent-GLIPCA」は特許文書構造を考慮したグラフベース手法を使用し，精度よく重要技術要素を抽出したうえで時系列連動性を分析する手法である．評価実験においては，特許文書からの重要技術語抽出およびトレンドのクラスタリング性能が比較手法よりも優れていることを示した．また，携帯電話に関する技術分野に適用したところ，ベース技術およびベース技術と異なるトレンドを持つ技術群を特定できることを確認した．

1　はじめに

特許情報に基づいてどのような技術がどのように時系列で変化しているのか，その技術トレンドを分析することは新技術の発見や技術動向分析に役立ち [1]，R&D 戦略策定に有益である．また，特許情報は M&A をはじめとする投資戦略策定の際に重要となる企業価値の評価に応用できることも示唆されている [2]．これらのことから，技術トレンドを分析する手法としての主流は特許文書中の引用情報に基づくものである．特に HITS[3] や PageRank[4] のように引用ネットワーク構造を評価する手法を用いて技術トレンドを分析する手法が確立されており [5]，特許中の技術トレンドと無形資産価格との関連を調べる研究がこれまでに行われている [6]．しかしながら，引用情報をベースとする技術トレンド手法の欠点はタイムラグが大きくリアルタイムでの分析を行うのが難しい点と具体的な技術内容はコンテンツベースの手法と組み合わせて分析する必要がある点があげられる．この課題を補うためには，具体的な技術要素を特許文書中から抽出してその時系列推移を分析する手法との併用が重要となる．特許や技術要素の技術的な推移に基づく分析手法としては [7] や [8] が提案されている．しかしながら，これらの手法には静的に特許文書をクラスタリングする手法 [7] は文書単位でのクラスタリングであるため粒度の細かな技術要素の推移をおうことはできないという問題がある．また技術要素をベースとする時系列推移を分析する手法 [7] も技術要素は特にフィルタリングせずに抽出しており技術語の抽出精度に課題が存在した．さらに PCA を実行して静的に技術的な語をまとめ上げた後に時系列推移を分析する手法であるため，時間変化の相関を見ておらず技術要素の連動性までは分析できなった．本研究はこれらの課題を解決するため，精度よく特許文書中から発明の主題となっている重要な技術要素のみを抽出したうえで，疑似相関を排して時系列連動性を分析できる手法として Patent-GLIPICA を提案する．

2　提案手法

提案手法の概要を図 1 に示す．本手法では，抽出された各クラスタの Purity の平均で評価する．Purity は，クラスタリングアルゴリズムの性能を評価するために一般的に使用される尺度である。Purity が高いほど，優れたクラスタリングアルゴリズムを示す．クラスタ C_i の Purity である $P(C_i)$ は以

下の式で計算される．

$$\mathrm{P}(\mathrm{C}_i) = \frac{1}{n_i} \max_k n_{i,k} \tag{1}$$

ここで，n_i はクラスタ C_i 内の単語数であり，$n_{i,k}$ はクラスタ C_i 内のクラス k に属する単語数である．

クラスタの抽出方法は，特許からの重要技術語抽出 2.1，重要技術語の時系列トレンドクラスタリング 2.2 からなる．

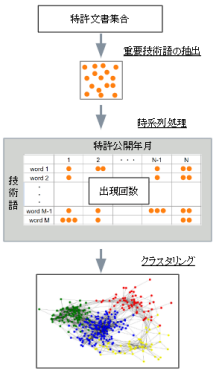


図 1: 提案手法の概要

2.1 重要技術語の抽出

特許からの重要技術語抽出は，特許文書構造を考慮したグラフベース手法を使用して抽出する．本手法は，特許文書内の項目間の意味関係を有向グラフとして表現した，教師なしのグラフベース手法である．構築した有向グラフの例を図 2 に示す．

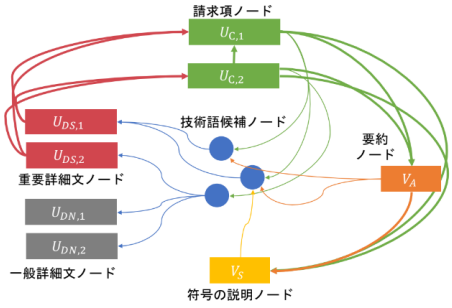


図 2: 特許から構築される有向グラフ

請求項、要約、符号の説明といった項目に含まれる技術語候補は重要である。そのような技術語候補の PageRank スコアが高くなるように有向グラフを構築している。

技術語候補 T_i は，**(形容詞)*(名詞)+**の正規表現パターンに一致するフレーズとし，要約，特許請求の範囲，明細書から抽出する．重要詳細文は坂地らの Cross-Bootstrapping 法 [9] にを用いて抽出した．重要詳細文の例を図 3 に示す．

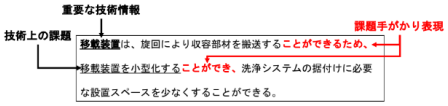


図 3: 重要詳細文の例

最後に各ノードの PageRank スコアを計算し，スコア上位 N 件の技術語候補を特許文書における重要技術語として抽出する．

2.2 時系列トレンドクラスタリング

本手法の重要技術語のトレンドクラスタ抽出は，原川らのトレンドクラスタリング手法 GLIPCA[10] を改良したものであり，G-GLIPCA と表記する．入力するデータは，平滑化，標準化を施した時系列データである．

GLIPCA はまず，グラフィカル・ラッソ・アルゴリズム [11] を使用する．このアルゴリズムは，直接相関を持つ単語のみを接続する部分相関ネットワークの構築を可能にする．具体的には，グラフィカル・ラッソ・アルゴリズムは共分散行列 S を用いて，精度行列のスパース推定を次の式で計算することによって得られる．

$$\hat{\Sigma}^{-1} = \underset{\Sigma^{-1}}{\operatorname{argmax}} \left\{ \ln\det(\Sigma^{-1}) - \mathrm{tr}(S\Sigma^{-1}) - \rho\|\Sigma^{-1}\|_{\ell_1} \right\} \tag{2}$$

次に精度行列 Σ^{-1}から偏相関行列を求め，第一主成分を取得する．第一主成分を取得は具体的には次の式を満たす $u = [u_1, u_2, \ldots, u_M]^T$を計算する．

$$L^T u = \lambda u \tag{3}$$

λ は第 1 主成分，u は対応する固有ベクトルである．各固有ベクトルの値は，各単語に対応する．固有ベクトルの値が一定以上の場合，その単語はクラスタ要素として抽出される． クラスタに含まれた単語を偏相関ネットワークから除いてから，式 (3) に戻って次のクラスタを取得する．この手順を繰り返し実行することで、重複のないクラスタを得ることができる。GLIPCA の流れを図 4 に示す．

クラスタリング結果は，モジュール性 Q によって定量的に評価され，Q が最大となる時系列データの平滑化期間 B とグラフィカルラッソ正則化強さ ρ が一意決定される．モジュール性 Q とは，ネットワークやグラフのクラスタリングの評価指標の一つ

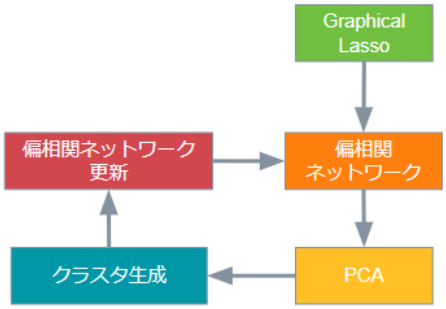


図 4: GLIPCA 概要

であり，ネットワーク内のノードがクラスタにどれだけ密に結合し，クラスタ間がどれだけ疎に結合しているかを評価するものである．

GLIPCA は入力時の単語の時系列情報に基づいて作成された偏相関行列から，クラスタを取得した後，ネットワーク L を更新することで繰り返しクラスタを取得している．しかし，更新されたネットワークは入力時の単語の時系列情報に基づいたものではない．そのため，G-GLIPCA では，クラスタ取得後，クラスタに含まれた単語を削除し，式（2）から実行し直す．G-GLIPCA のクラスタリングの流れを図 5 に示す．

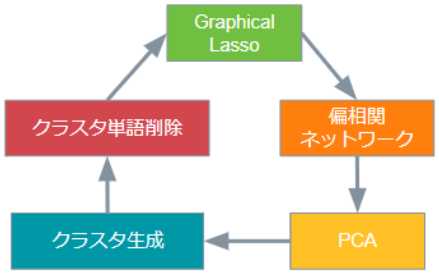


図 5: G-GLIPCA 概要

3 実験

3.1 重要技術語の抽出

本研究では，重要技術語の抽出性能を評価するために 1993-2002 年の IPC コードのセクション A H に属する特許をセクションごとに 10 件ずつランダムサンプリングした．また，評価のために，特許分析を専門とするアノテータ複数名で重要技術語のアノテーションを行った．なお，化学・冶金分野のセクション C の特許はデータセットから除外した．化学・冶金分野の特許は，発明において重要な要素が化学式で表現されることが多いためである．

提案手法の有効性を検証するために，複数の教師なしキーワード抽出手法との間で重要技術語抽出の性能の比較を行った．比較手法として，統計的手法の TF-IDF，グラフベース手法の TextRank[12]，PositionRank[13] を選択した．抽出性能の評価指標には Precision(P)，Recall(R)，F 値 (F1) を用いた．

3.2 時系列トレンドクラスタリング

対象とした特許は，日本国特許庁が定める FI ターム H04M1/02C に分類されるもので，1996-2002 年までの 1990 件を対象とした．

G-GLIPCA の有効性を検証するために，GLIPCA との比較と行った．特許 1 件から抽出する重要技術語の個数は 5 個とした．重要技術語は特許公開年月に基づいて出現回数とカウントし，期間中 1 回しか出現しない単語は削除した．平滑化期間 B は 1-12 ヶ月，グラフィカルラッソの正則化強さ ρ は 0.1-1 の 0.1 間隔の間で最適化パラメータを決定した．抽出性能の評価指標には各クラスタの平均 Purity を用いた．

4 結果

4.1 重要技術語抽出

評価結果を表 1 に示す．全セクションでの評価結果を見ると，提案手法は F 値が 67.93%と最も高かった．これは，比較手法の中で最も抽出性能が高い PositionRank よりも F 値が 49.85 ポイント高い．さらに，各セクションごとの結果でも，提案手法が一貫して最も高い抽出性能を示している．

表 1: 重要技術語抽出の評価結果

	提案手法			TF-IDF			TextRank			PositionRank		
Section	P(%)	R(%)	F_1(%)	P(%)	R(%)	F_1(%)	P(%)	R(%)	F_1(%)	P(%)	R(%)	F_1(%)
A	**58.00**	**74.36**	**65.17**	4.00	5.13	4.49	4.00	5.13	4.49	8.00	10.26	8.99
B	**64.00**	**78.05**	**70.33**	20.00	24.39	21.98	6.00	7.32	6.59	20.00	24.39	21.98
D	**58.00**	**52.73**	**55.24**	26.00	23.64	24.76	18.00	16.36	17.14	24.00	21.82	22.86
E	**76.00**	**80.85**	**78.35**	16.00	17.02	16.49	8.00	8.51	8.25	20.00	21.28	20.62
F	**72.00**	**67.92**	**69.90**	16.00	15.09	15.53	10.00	9.43	9.71	20.00	18.87	19.42
G	**74.00**	**75.51**	**74.75**	12.00	12.24	12.12	12.00	12.24	12.12	16.00	16.33	16.16
H	**64.00**	**61.54**	**62.75**	14.00	13.46	13.73	10.00	9.62	9.80	16.00	15.38	15.69
All	**66.57**	**69.35**	**67.93**	15.43	16.07	15.74	9.71	10.12	9.91	17.71	18.45	18.08

4.2 時系列トレンドクラスタリング

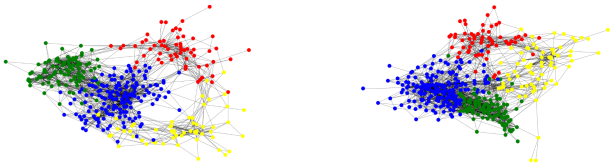
評価結果を表 2 に示す．G-GLIPCA では Purity が 0.704 であり，GLIPCA よりも Purity が 0.033 ポイント高い．

次に，各手法によって抽出されたクラスタのを図 6 に示す．抽出されたクラスタは 4 つであり，抽出された順に第 1,2,3,4 クラスタとし，それぞれ青，

表 2: クラスタリングの評価結果

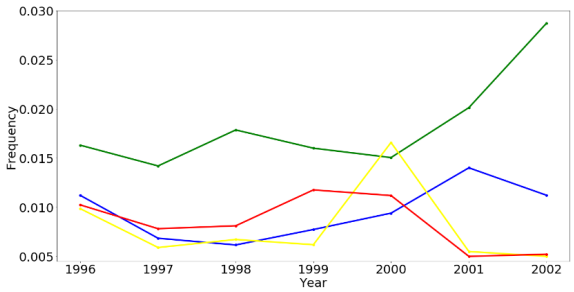
	G-GLIPCA	GLIPCA
Purity	0.704	0.671

緑，黄，赤で表示している．さらに各クラスタのトレンドを図 7 に示す．縦軸は特許 1 件あたりに各クラスタの何%の単語が含まれるかを年ごとに集計したものを示す．横軸は対象期間である．

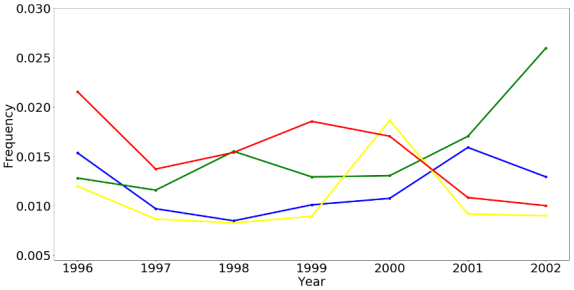


(a) G-GLIPCA　　(b) GLIPCA

図 6: 抽出されたクラスタ



(a) G-GLIPCA



(b) GLIPCA

図 7: クラスタのトレンド

5　考察

5.1　重要技術語抽出

比較手法のスコアが低い要因として，技術語候補スコアの算出方法が挙げられる．提案手法と同様にしてスコア算出の単位を単語から技術語候補に変更すると，ほとんどの技術語候補の出現頻度が非常に低くなる可能性があるため，TF-IDF のような頻度に基づく手法は，技術語候補のスコアがうまく計算できなくなる恐れがある．一方で，TextRank や PositionRank といった従来のグラフベースの手法では，ある範囲内での候補単語・フレーズの共起に基づいてエッジを設けている．このことは，多くの修飾語句を用いて用語の意味を限定する記述をしている特許文書 [14] において技術語周辺の修飾語句の影響が強く現れてしまう可能性がある．従来のグラフベースの手法は，これらの理由から，語の統計量に依存している従来手法のアプローチでは，重要技術語抽出は困難であることが示唆される．一方，提案手法は，特許文書の意味的な構造に着目することで上記の問題を回避しており，結果として最も高い抽出性能を示したと考えられる．

5.2　クラスタリング

比較手法である GLIPCA でスコアが低い要因として，2 個目以降のクラスタの算出方法が挙げられる．GLIPCA では，クラスタに所属している単語が除かれることによる統計情報の変化を加味せずに，次のクラスタを生成してしまうことがスコアの低下につながったと考えられる．一方で，G-GLIPCA はクラスタ生成ごとに変わる統計情報に基づいて，毎回偏相関行列を計算し直すため，上記の問題を解決しており，GLIPCA の Purity を上回ったと考えられる．

クラスタごとに含まれている単語は，G-GLIPCA，GLIPCA 両方で，第 1 クラスタにベース技術，第 2 クラスタに折りたたみ式携帯技術，第 3，4 クラスタは周辺機器のような，ベース技術とそれとはかけはなれたトレンドをもつ群がそれぞれ抽出された．第 1 クラスタはベースとなる技術ということもあり特に特徴のないトレンドとなった．また，第 2 クラスタは上昇トレンドを描いた．これは当時折りたたみ式携帯電話が主流になった時期であり，開発が著しく増加した背景を示していると考えられる．第 3 は低い位置，第 4 は下降傾向にあり，当時，開発されつくして低迷傾向にあった技術群が抽出されたと考えられる．

6　まとめと今後の展望

本研究では，特許情報に基づく技術トレンド分析手法として，「Patent-GLIPCA」を開発した．比較実験の結果，従来手法と比べて高い性能を示すことを確認した．また，評価実験によりトレンドの異なる技術群を特定できることを確認した．今後は，さらなる手法の改良と企業価値評価や技術予測などのタスクに応用することを行っていく．

謝辞

本研究は JSPS 科研費 JP19K12116 の助成を受けたものです.

参考文献

[1] Péter Bruck, István Réthy, Judit Szente, Jan Tobochnik, and Péter Érdi. Recognition of emerging technology trends: class-selective study of citations in the U.S. Patent Citation Network. **Scientometrics**, Vol. 107, No. 3, pp. 1465–1475, June 2016. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer Netherlands.

[2] Hirofumi Nonaka, Daiki Kubo, Toru Hiraoka Kimura, Takahisa Ota, Shigeru Masuyama, et al. Correlation analysis between financial data and patent score based on hits algorithm. In **2014 IEEE International Technology Management Conference**, pp. 1–4. IEEE, 2014.

[3] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, Vol. 46, No. 5, pp. 604–632, September 1999.

[4] The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, Vol. 30, No. 1-7, pp. 107–117, April 1998. Publisher: Elsevier.

[5] Kensei Nakai, Hirofumi Nonaka, Asahi Hentona, Yuki Kanai, Takeshi Sakumoto, Shotaro Kataoka, Elisa Claire Alemán Carreón, and Toru Hiraoka. Community Detection and Growth Potential Prediction Using the Stochastic Block Model and the Long Short-term Memory from Patent Citation Networks, April 2019.

[6] Yuta Yamamoto, Asahi Hentona, Koji Marusaki, Kohei Watabe, Seiya Kawano, Tokimasa Goto, Yutaka Hada, Kazuhisa Fukuzawa, and Hirofumi Nonaka. Development of the patent values evaluation method considering growth of technical community. In **2021 IEEE Symposium Series on Computational Intelligence (SSCI)**, pp. 1–6, February 2021.

[7] Patent text mining based hydrogen energy technology evolution path identification. **International Journal of Hydrogen Energy**, Vol. 49, pp. 699–710, January 2024. Publisher: Pergamon.

[8] **Technology Roadmapping of Emerging Technologies: Scientometrics and Time Series Approach**.

[9] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁. Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法. 電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp. 742–755, 2010.

[10] Ryosuke Harakawa, Tsutomu Ito, and Masahiro Iwahashi. Trend clustering from COVID-19 tweets using graphical lasso-guided iterative principal component analysis. **Sci Rep**, pp. 5709–5709, 2022.

[11] Friedman J, Hastie T, and Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. **Biostatistics (Oxford, England)**, Vol. 9, No. 3, July 2008. Publisher: Biostatistics.

[12] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Text. pp. 404–411, July 2004.

[13] Corina Florescu and Cornelia Caragea. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. pp. 1105–1115, July 2017.

[14] 潮田明. 特許分における長い名詞句表現の自動解析について. **In Japio YEAR BOOK**, Vol. 6, pp. 274–279, 2012.