特許構成要件の置換による新たな特許検索・技術創出モデルの基礎的検討

代表研究者	野 中 尋 史	長岡技術科学大学	情報・経営システム工学専攻 准教授
共同研究者	平 岡 透	長崎県立大学	情報システム学部 教授

1 背景・目的

情報通信分野における自然言語処理/テキストマイニングの研究が隆盛を迎えている.このような中で,Web文書をはじめとする一般になじみが深い電子文書に関しては検索技術が確立されつつある.

一方，特許権に基づく発明の保護と活用は，知財立国を国家戦略の基本に掲げる日本において重要な課題となっている．そのような背景のもと特許文書を対象とする特許情報処理に関して，現在，様々な内容の研究が実施されている．中でも特許検索に関わる技術はその重要性にも関わらず，従来研究の性能は MAP で10-20%程度に留まっており，実用化のためには大幅な性能改善が求められている．特許の審査においては発明内容を構成要件と呼ばれる技術構成要素に分解した上で，全ての構成要件が一致した従来特許（技術）が存在するかどうか（新規性），もしくは，別の従来特許（技術）の構成要件を組み合わせ・置換することにより容易に実現できるかどうか（進歩性）の観点で行われる．しかしながら，従来の特許検索に関わる技術はこのような審査プロセスを意識したものではなく，Web 検索で使用される手法の応用が主であった．特許検索の研究においては，後述の通り， TF-IDF 等をベースとする Web 検索の応用に留まっており，また，その性能も MAP 値が 0.5 を大菊下回るなど実用的なレベルではなかった([1]-[6])．性能向上のためには，特許審査プロセスにおける新規性・進歩性を担保する手法の開発が求められていた．すなわち，発明を構成要件に分割したうえで，完全な構成要件の一致（新規性）と一部を他特許と置換した構成要件との一致（進歩性）を判定する必要がある．文書全体の内容に基づく類似度をベースとする Web 検索の手法では，構成要件毎の類似度を判定することは不可能である．また，単純な審査官引用をベースとする手法でも，置換操作などを加味することは難しい．新規性・進歩性に関しては，技術内容とその効果から解釈される構成要件へ発明を分割し，組み合わせ爆発を起こすことなく構成要件の置換を行う技術の確立が必要である．これらの技術が確立されると特許検索のみならず，これまでにない構成要件からなる新規技術案の提案にもつながる．このため，審査プロセスを意識した検索手法の確立が求められている．

以上を整理すると特許検索手法として

(1) 技術内容を構成要件に分解すること

(2) 引用ネットワーク（特許審査における引用関係をリンクと見立てたネットワーク）クラスタリングにより組み合わせ爆発を防ぎながら他特許の構成要件へ当該技術の構成要件を置換すること

(3)構成要件の置換可能性を判定すること（可能であれば特許要件は拒絶され，困難であれば新規性・進歩性を満たす発明として成立する）

の 3 つの要件を満たす手法の開発が求められる．

これらの課題を解決した手法が確立されると特許検索の性能が格段に向上することが期待されることはもとより，論文検索等へも拡張することで幅広い技術探索に応用できる．さらに，ベース特許・技術に対して現状では構成要件を置換困難なものに置換した（＝新規性・進歩性を満たす）発明案を列挙することで新しい技術的アイディアを提案することにもつながる．

そのような中で申請者らは，特許で構成要件に関連する特許文書からの技術・効果に関する表現抽出や，進歩性判定の組み合わせ判定の基礎となる引用ネットワークのクラスタリングとその応用に関する研究([6],[7])を進めていた.引用ネットワークのクラスタリングについても正解率が 0.8 以上を達成している．特許文書からの技術・効果に関する表現抽出に関しては，ブートストラップ法と文法パターンを利用した手法（[8],[9])により正解率 0.9 以上を達成している．そこで，これらの技術に基づき，特許審査プロセスに基づく技術探索，および，新規技術案の提案を行う手法のうち上記(1)「技術内容を構成要件に分解すること」と(2)「引用ネットワーク（特許審査における引用関係をリンクと見立てたネットワーク）クラスタリングにより組み合わせ爆発を防ぎながら他特許の構成要件へ当該技術の構成要件を置換すること」について基礎的検討を行うことが本申請の目的である．

公益財団法人電気通信普及財団
研究調査助成報告書 第35号 2020年度

2 研究手法

2-1 研究全体の概要

以下の 2 つに関する手法の確立を行った.

(a)構成要件の抽出技術の開発

特許における構成要件は，技術内容とその効果の対によって表現される．これまでの申請者らの研究[8]および[9]により，効果に相当する語は特定の手がかり語（「できる」等）と係り受けの関係になっていることが分かっており，手がかり語を自動的に獲得できるブートストラップ法により獲得済みである．しかし，今回は構成要件としての技術語を抽出した上で，効果語と技術語の対が必要であるため，重要技術語を抽出する新たな手法の開発が求められる．

そこで，特許文書全体の構造情報とその意味関係をグラフで表現したグラフベースの教師なし重要技術語抽出手法の開発を行った．具体的には，特許の各項目と技術語候補をノード，それらノード間の意味関係をエッジとしてグラフを構築し，グラフベースのランキングアルゴリズムである PageRank を適用することで，技術語候補のスコアリングを行った．本提案手法は教師なし手法であるため，非常にコストがかかる特許文書のアノテーションを行う必要はない.

(b)審査官引用ネットワークに基づくネットワーククラスタリングの確立

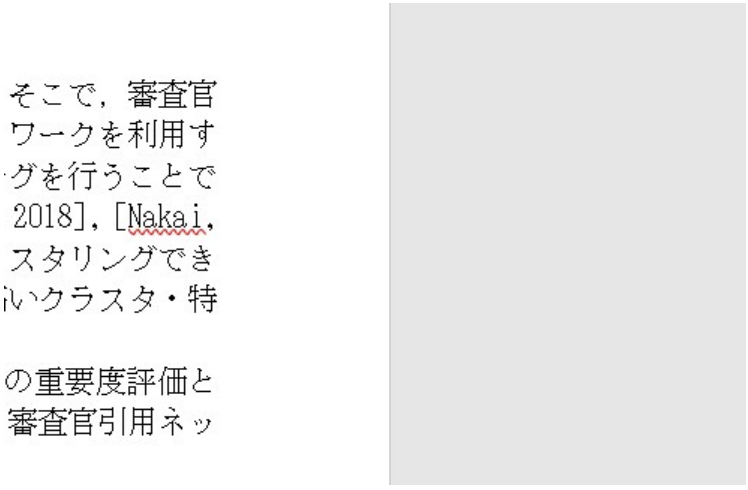
構成要件置換を行う対象は組み合わせ爆発の観点から類似特許群に絞り込む必要がある．そこで，審査官が審査した特許とその審査の時に引用した特許の関係をリンクと見たてた審査官引用ネットワークを利用する．審査官引用ネットワークに対してネットワーク埋め込みベクトルを構築しクラスタリングを行うことで類似特許群を同定する．これまでの申請者らの研究([6],[7])により Node2vec を用いた場合，正解率 0.8 の性能でクラスタリングできることが分かっている．ただし，各クラスタ・特許の重要性は異なるため，より重要度の高いクラスタ・特許を検索対象とするほうが検索効率がよいと推測される．

そこで，本研究では，引用成長性の予測モデルによる分野・コミュニティ（クラスタ）の重要度評価と PageRank に基づく個別特許の重要性評価を掛け合わせた特許スコアリングモデルを開発し，審査官引用ネットワークに基づくネットワーククラスタに基づく対象絞込の精緻化を行った.

2-2 構成要件の抽出技術の開発

本項目では，特許文書の構造を使用する．特許文書の構造を表 1 に示す．表 1 に示すような書類と主要項目で特許文書の構造は規定される.

表 1．特許文書の構造



本項目では，特許文書の構造を使用する．特許文書の構造を表 1 に示す．表 1 に示すような書類と主要項

目で特許文書の構造は規定される．

これら構造における各項目と構成要件（重要技術語）の意味的関係性に着目し，本研究では，以下の図で表現されるようなグラフである特許構造グラフを規定した．

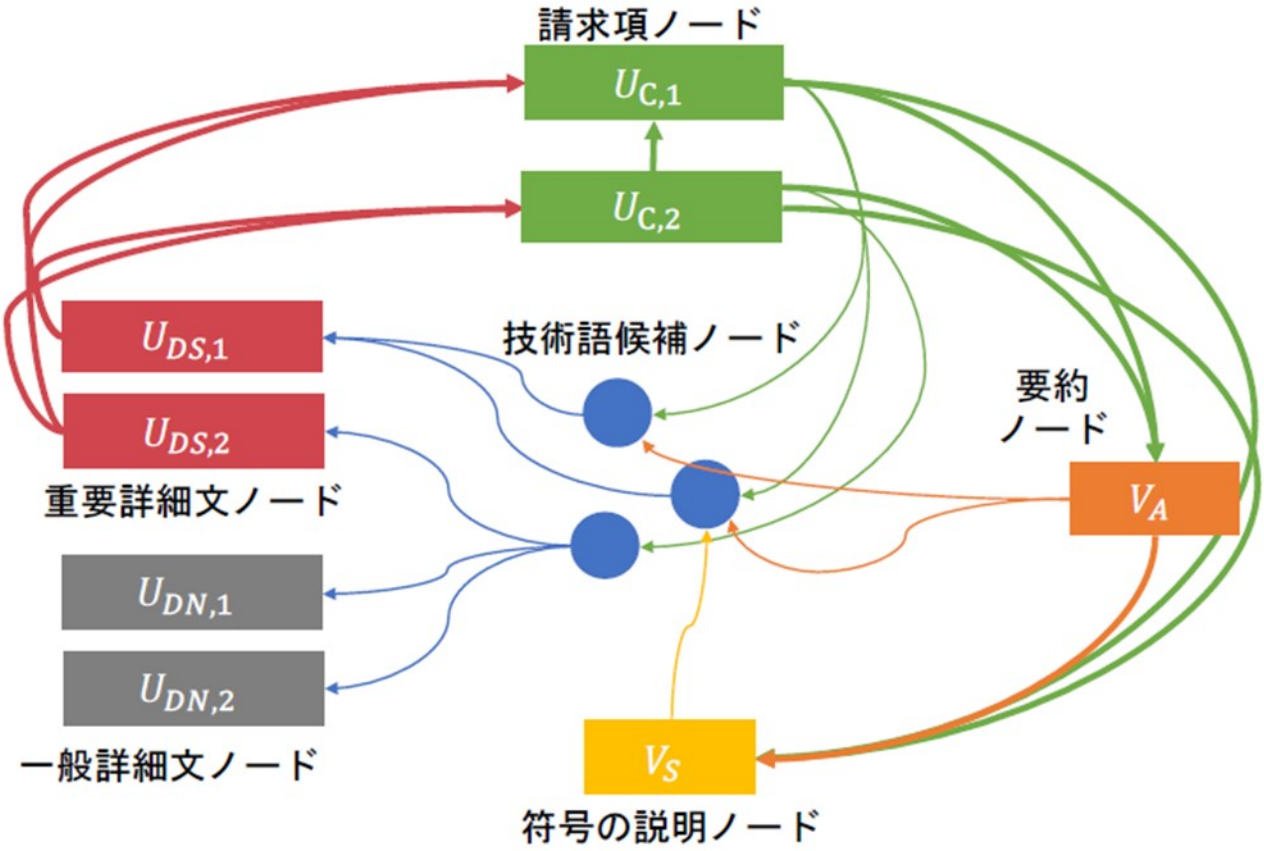


図 1. 構成要件の抽出のための特許構造グラフネットワーク

本グラフは，特許文書内の項目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．有向グラフは以下の仮定に基づいて特許文書から構築している．この有向グラフに後述する PageRank 様のスコアリング手法を適用し，高いスコアとなる技術語が構成要件として抽出される．

・発明の技術的特徴の要約度合は，請求項，「要約」，「符号の説明」の順に高くなっていく．従って，各項目の重要度は，請求項<「要約」<「符号の説明」となる．

・ 独立請求項は，従属請求項よりも重要である．

・ 明細書において，技術上の課題が含まれている文には，その課題を解決する重要な技術情報が含まれていることがあるため重要である．一方，その他の文には一般的な技術情報が多く含まれており，発明の技術的特徴の記載は少ない．

本手法は，以下の 3 段階のステップで重要技術語を抽出する．

(a)技術語候補となるフレーズを選択する，

(b)特許文書から有向グラフを構築する

(c)グラフベースのランキングアルゴリズムを適用し，技術語候補をスコアリングする

(a)について詳細に述べる．(a)における技術語候補 Ti は，(形容詞)*(名詞)+のパターンに一致するフレーズとする．技術語候補は，「要約」「特許請求の範囲」「発明の詳細な説明」「符号の説明」セクションから抽出する．また，フレーズに含まれる単語数に制限は設けない．キーワード抽出の先行研究では，候補フレーズは bi-gram にするなど，単語数に制限を設けていることがある．しかしながら，厳密な記述が求められ

る特許文書には，複雑で長い名詞句を多く含んでいるという特徴がある．そのため，技術語候補となる候補フレーズの単語数には制限を設けないこととした．なお，特許特有の「請求項」などのフレーズもノイズとして削除する．

次に(b)の詳細について説明する.前述の有向グラフにおいては重要詳細文と一般詳細文がノードとして存在する．まず，この二つについて説明する．明細書には，発明の背景や課題，解決手段，効果，実施例など，発明に関する具体的な内容が記述されているため，他の項目よりも技術情報が多く含まれている．特に，技術上の課題が含まれている箇所には，その課題を解決するための重要な技術情報が含まれていると考えられる．そこで，技術上の課題が含まれている詳細文を重要詳細文，その他の詳細文を一般詳細文と定義する．

本研究では，Cross-Bootstrapping 法[9] により自動的に取得された「ことで、」「ことが可能であり、」「ようにしたため、」といった課題手がかり表現（30 件）を含む詳細文を重要詳細文とした．続いて，ノードについて説明する．本研究では，技術語候補ノード集合，請求項ノード集合，重要詳細文ノード集合，一般詳細文ノード集合，「要約」セクションノード，「符号の説明」セクションノードからなる．前述した仮定に基づき有向グラフを構築する．有向グラフで工夫されている点として，重要詳細文に多く出現する語が構成要件となりやすい傾向を反映するため，重要詳細文はため込みノード（技術語のスコアが減る）にならないように，一方で一般詳細文のみに多く出てくるものは構成要件でない可能性が高まるため，一般詳細文はため込みノードになりやすいグラフにしている．また，独立請求項に出てくる表現が最重要であることも考慮し請求項間の有向グラフも規定している．

(c)については PageRank で使用されるスコアを適用する．これにより，有向グラフ上の重要ノードを特定することができる．

2-3 審査官引用ネットワークに基づくネットワーククラスタリングの確立

審査官引用ネットワークに基づくネットワーククラスタリングの確立のためには，クラスタリングした結果である各クラスタ・特許の重要性を評価することが課題となっている．そこで，以下の図 2 で示される手法の開発を行った．ネットワーククラスタリングしたのちに，各クラスタの重要性を引用成長性で計測し，さらに各特許の当該ネットワークにおける重要性を PageRank で測定し,掛け合わせることで特許の重要度を評価している．これにより，検索対象の絞り込みが容易となる．

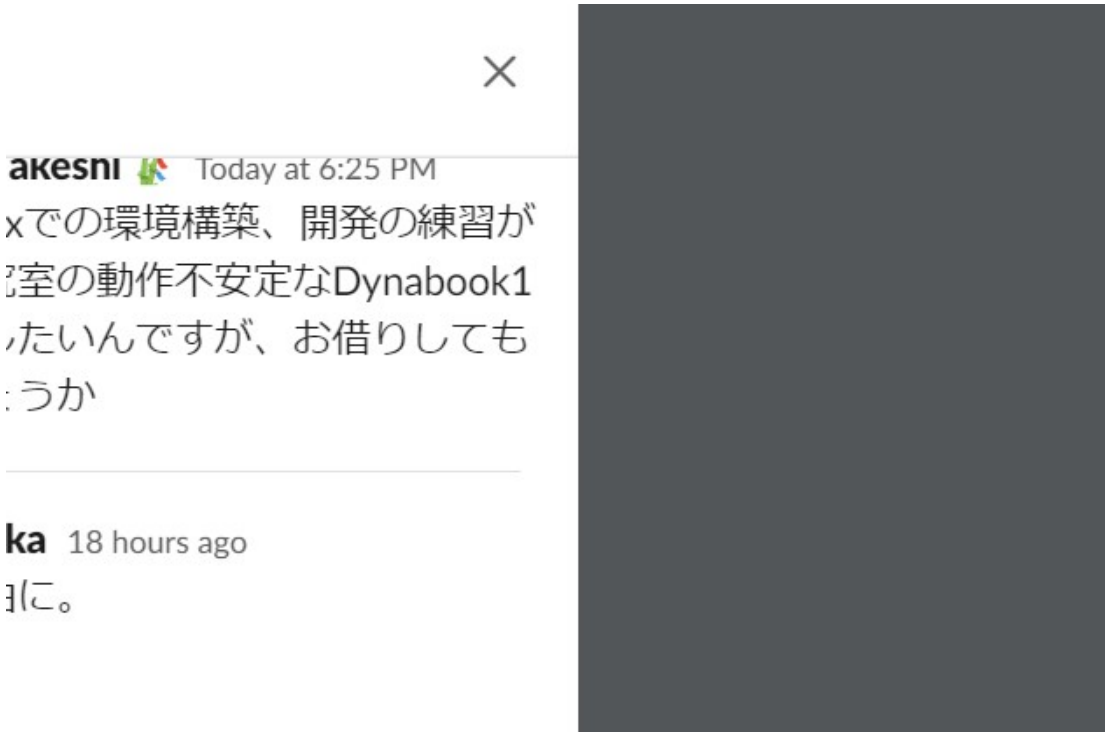


図 2．2-3 の研究概要

引用ネットワークのクラスタリングは，研究者グループが既に確立している Node2Vec 埋込に基づく x-means クラスタリングで行った.各クラスタの成長性予測は,ARIMA モデルによる時系列予測により行った.

特許単体の重要度は PageRank で計測した.

3 評価実験・考察

3-1 構成要件の抽出技術

本研究では，重要技術語の抽出性能を評価するためにオリジナルのデータセットを作成した．はじめに，日本語公開特許公報全文データ（期間：1993〜2002 年，文書数：3,496,252 件）から，国際特許分類のセクション 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．次に，サンプリングした各特許文書から，発明上特に重要と思われる技術語を，弁理士を含めた 3 名のアノテーターの合議により選択した．このとき，重要技術語は特許 1 件あたり 5 個を目安として選択した．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[10]，PositionRank[11] を選択した．TF-IDF の IDF スコアは，NTCIR-6 データセットからランダムサンプリングされた 10 万件の特許から計算したものを使用した．TextRank と PositionRank は，候補単語をノードとし，候補単語 wi と wj が前後 2 単語内で共起した場合にエッジを設けた無向グラフを構築した．これら 3 つの比較手法では，先行研究に従い，技術語候補を構成する単語のスコアの総和を，その技術語候補のスコアとした．PageRank アルゴリズムを用いているグラフ
ベースの手法（提案手法，TextRank，PositionRank）は，ダンピングファクター α を 0.85 に設定し，PageRank の反復計算を 100 ステップあるいは 1 ステップ前のスコア S との差が 0.001 より小さくなるまで実行した．
実験に際し，全ての手法に関して，特許 1 件から抽出する重要技術語の個数は 5 つとした．評価結果を以下の表示す．表より全セクションでの評価結果を見ると，比較手法の中で最も抽出性能が高い PositionRank よりも，提案手法の方が 48.94 ポイント F 値が高かった．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 2．評価結果

	提案手法			TF-IDF			TextRank			PositionRank		
Section	P	R	F1	P	R	F1	P	R	F1	P	R	F1
A	50.00	86.21	63.29	4.00	6.90	5.06	4.00	6.90	5.06	6.00	10.34	7.59
B	52.00	83.87	64.20	18.00	29.03	22.22	6.00	9.68	7.41	18.00	29.03	22.22
D	50.00	55.56	52.63	18.00	20.00	18.95	16.00	17.78	16.84	18.00	20.00	18.95
E	66.00	89.19	75.86	14.00	18.92	16.09	6.00	8.11	6.90	16.00	21.62	18.39
F	70.00	81.40	75.27	12.00	13.95	12.90	10.00	11.63	10.75	14.00	16.28	15.05
G	62.00	77.50	68.89	8.00	10.00	8.89	12.00	15.00	13.33	12.00	15.00	13.33
H	52.00	61.90	56.52	14.00	16.67	15.22	10.00	11.90	10.87	16.00	19.05	17.39
All	57.43	75.28	65.15	12.57	16.48	14.26	9.14	11.99	10.37	14.29	18.73	16.21

特許 10 件分の抽出結果をランダムサンプリングし，比較手法のエラーアナリシスを行った．各手法ごとに誤抽出の内容を見ると，技術語と思われるフレーズに不必要な単語が付いた技術語候補を誤抽出しているケースが多数確認できた．例えば TF-IDF では，1 件の特許文書から“各シェル”，“シェル上”といった，“シェル”が含まれた技術語候補を複数個誤抽出していた．このタイプの誤抽出が，TF-IDF で 45 件中 24 件，TextRank で 44 件中 17 件，PositionRank で 44 件中 16 件存在した．
このような誤抽出が生じる要因として，比較手法の技術語候補スコアの算出方法が考えられる．比較手法では，単語ごとにスコアを求めており，技術語候補のスコアは構成単語のスコアの総和としている．しかしながら，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．そのため，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，スコア算出の単位を単語から技術語候補に変更したとしても顕著な効果は現れないと考えられる．その理由として，厳密さが求められる特許文書では，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ことが挙げられる．従来のグラフベースの手法は，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ため，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．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と考えられる．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いるため，結果として最も高い抽出性能を示したと考えられる．

3-2 審査官引用ネットワークに基づくネットワーククラスタリングの確立

本研究では，知的財産研究所の特許データベースに格納されている 600 万件のデータを対象とした．ARIMA での予測については，1980/1/1-2000/12/31 までの出願特許を対象にモデル構築を行い，2001 年度における引用数の予測を行った．予測の評価は SMAPE で行った．

IPC の大分類ごとの SMAPE 値は以下の図のとおりである．比較手法として LSTM も載せている．結果から不確実性の高い技術開発分野の時系列予測においても多くの分野で 20%より小さい値となっており実用レベルであることが分かった．また，PageRank と掛け合わせた最終的なスコアに関して東証 1 部上場の電気機器セクターが出願した特許のスコア上位 10 件を調査したところ，広告配信など ICT 分野の重要特許が抽出できていることが分かった．

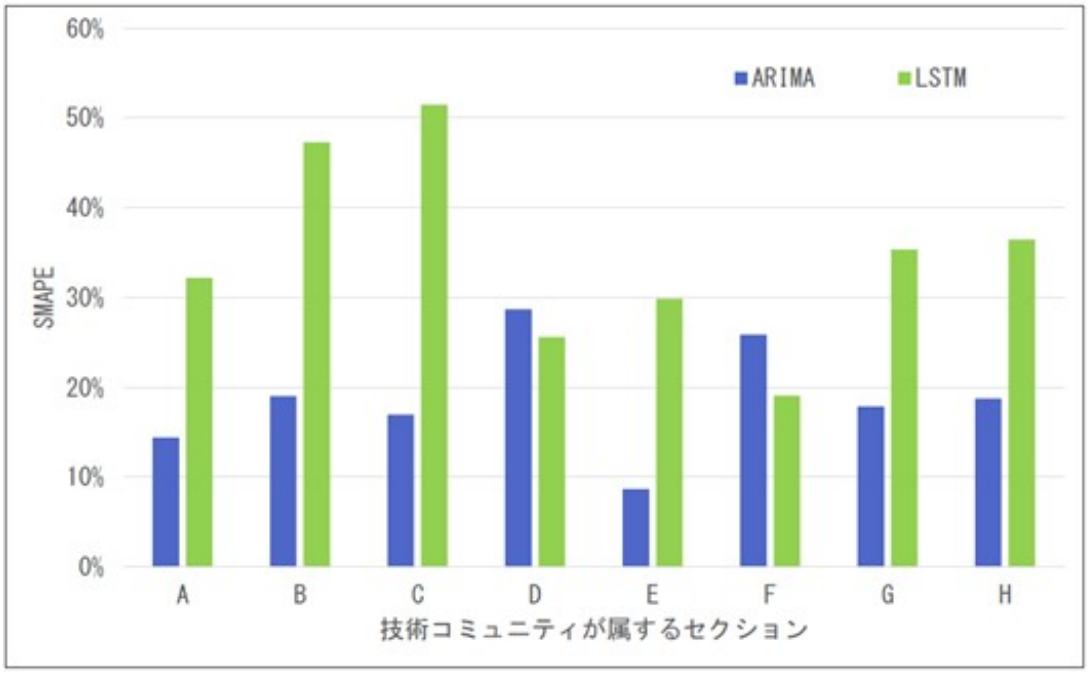


図 3. IPC セクションごとの SMAPE の結果．ARIMA と比較手法として LSTM の結果も載せている．

4 結論

特許審査プロセスに基づく技術探索，および，新規技術案の提案を行う手法の基礎的検討として「技術内容を構成要件に分解すること」を実現するためのグラフベース手法を提案し，さらに「引用ネットワーク（特許審査における引用関係をリンクと見立てたネットワーク）クラスタリングにより組み合わせ爆発を防ぎながら他特許の構成要件へ当該技術の構成要件を置換すること」を実現するための引用成長性予測と PageRank を組み合わせたスコアリング手法の開発を行った．評価実験の結果，前者は既存手法を凌駕する手法を確立でき，後者についても重要クラスタ・特許のスコアリングが可能であることを示した．今後は構成要件の置換可能性の手法の確立を行い，特許審査プロセスに基づく技術探索，および，新規技術案の提案を行う手法の実用化を目指す．

【参考文献】

[1] Hisao Mase, et al., Proc of NTCIR-6, 2006.

[2] Masaki Rikitoku et al., Proc of NTCIR-6, 2006.

[3] Hidetsugu Nanba et al., Proc of NTCIR-6, 2006.

[4] P. Mahdabi and F. Crestani, Proc of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1659–1668, ACM, 2014.

[5] W. Tannebaum, et al., Multimodality, and Interaction, pp. 300–305, Springer, 2015.

[6] Asahi Hentona, Hirofumi Nonaka, Kensei Nakai, Takeshi Sakumoto, Shotaro Kataoka, Elisa Claire Aleman Carreon, Hugo Alberto Mendoza Espana, Toru Hiraoka, Masaharu Hirota, Community Detection and Growth Potential Prediction from Patent Citation Networks, Proc. of 2018 ACM International Conference on Management of Digital EcoSystems(MEDES), 2018.

[7] Kensei Nakai, Hirofumi Nonaka, Asahi Hentona, Yuki Kanai, Takeshi Sakumoto, Shotaro Kataoka, Elisa Claire Aleman Carreon, Toru Hiraoka, Community Detection and Growth Potential Prediction Using the Stochastic Block Model and the Long Short-term Memory from Patent Citation Networks, Proc. of 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEE IEEM), Singapore, 2018.

[8] Hirofumi Nonaka, Akio Kobayashi, Hiroki Sakaji, Yusuke Suzuki, Hiroyuki Sakai, Shigeru Masuyama, Extraction of Effect and Technology Terms from a Patent Document, Journal of Japan Industrial Management Association, vol. 63, no. 2E, pp.105-111, 2012.

[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, 情報・システム J93-D(6), pp. 742-755, 2010.

[10] Mihalcea, R. and Tarau, P.: TextRank: Bringing Order into Text, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411, Barcelona, Spain (2004), Association for Computational Linguistics.

[11] Florescu, C. and Caragea, C.: PositionRank:An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1105–1115, Vancouver, Canada (2017), Association for Computational Linguistics.

〈発 表 資 料〉

題　　名	掲載誌・学会名等	発表年月
技術コミュニティの成長性を加味した特許価値評価手法の開発	人工知能学会全国大会2020(JSAI2020)	2020 年 6 月