論　文

特許引用をベースとする特許価値評価手法に関するレビュー

野　中　尋　史

1．はじめに

特許は企業の発明を守り、経済的利益をもたらすことができる[1)]。さらに特許は、様々な分野で出願され、他の統計指標と異なり細かな技術分野に焦点をあてた分析も可能であり、アクセスもしやすい特徴を持つ[2)]。このような特徴から特許データは、製品の開発状況、市場競争、あるいは技術マネジメントを分析するための重要な指標として幅広く利用されている[2)-5)]。特許情報を抽出するために、特許に特化したテキストマイニング手法に関する研究も増えている。内田ら[6)]は、概念に基づくベクトル空間モデルを用いて、パテントマップを自動生成する方法を開発した。石川ら[7)]は、「ことより」を手掛かりに、「コーティング材」のような技術用語と「耐衝撃性」のような効果用語を特許文書から抽出することを試みた。野中ら[8)]は、エントロピーに基づく手法と文法パターンを用いて、技術用語と効果用語の抽出手法を提案した。このように特許情報の抽出手法は現在までに多数提案されている。一方で特許情報の解析の観点としては他に特許価値評価に関する研究が存在する。この分野で特に利用されているのは特許文書の引用数である。インパクトファクターのような引用文献の数は、同じ技術分野における他の競合他社の注目度を示す。したがって、多くの文献に引用されている特許は高得点とみなすことができる。実際にHallら[9)]は、特許引用度が市場価値に有意に影響することを見出している。しかし、このような単純な引用のカウントには限界がある。引用には、引用の引用のような連鎖的な引用構造が存在する。ソーシャルネットワーク分析ではこのようなリンクの連鎖構造を評価すること重要とされている。このため特許においてもPagerankやHITSのようなリンク構造評価手法の適用[10)]が必要とされていた。こうしたなかでソーシャルネットワーク分析での研究に基づいてLukachら[11)]は、特許のPageRankスコアで重要度を計算することを提案した。この研究では、

PageRankの重要度の重み付けは、後方引用や前方引用の数による重み付けとは異なっていることを明らかにしている。しかしながら、Lukachらの研究ではPagerankのような引用ネットワークが企業価値と関係しているかどうかまでは明らかにできていなかった。そこで、本論文では引用ネットワークのリンク構造を評価し企業価値や技術分野の成長性と関連付けて特許の重要性を算出する手法のレビューを行う。具体的には引用ネットワークのリンク構造が特許の価値評価において重要なことを示唆したNonakaらが行った研究[12] と引用リンクコミュニティの成長性を評価するHentonaらの研究[13]の紹介を行う。

2．引用ネットワークのリンク構造評価

本章では、引用ネットワークのリンク構造により特許を評価したNonakaらの研究[12] の紹介を行う[12]。**図1**に分析の概要を示す。まず、日本の特許文献のHITSオーソリティスコアを算出する。これらの特許文書は、1996年から2006年まで適用されたものである。特許文書の総数は、3,875,711件である。次に、特許出願人のHITSオーソリティスコアの総和として、特許出願人ポートフォリオスコアを算出した。最後に、株式データと上記スコアの相関分析を行った。この研究では、東京証券取引所に上場している企業に焦点を当てている。

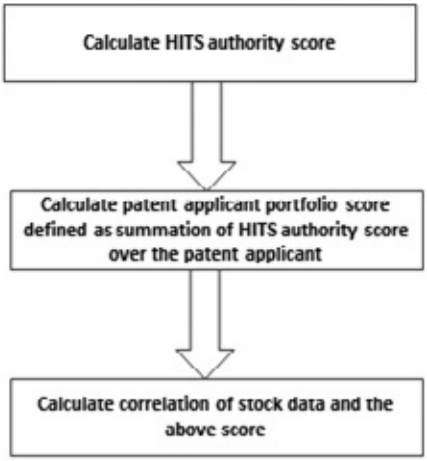

図1．研究の流れ

ここでHITSアルゴリズムについて説明する。このアルゴリズムはJon Kleinbergによって初めて提案された[10]。HITS（Hyperlink Induced Topic Distillation）アルゴリズムの背後にある考え方は、権威とハブが相互に補強し合うというものである。ページの権威の重みは、そのページを指すハブの重みの合計として計算され、ハブの重みは、そのページが指す権威の重みの合計として計算される。言い換えれば、良いハブは他の多くのページを指しているページを表し、良いオーソリティは多くの異なるハブによってリンクされているページを表す。以下にHITSの理論式を記す。Authorityベクトルおよびハブの重みベクトルの計算は、反復して行われる。k回目の反復における権威の重みベクトルは(1)のように定式化される。また、k回目の反復におけるハブの重みベクトルは、(2)のように定義される。ここで、Aは隣接行列である。

$$\vec{a_k} = A^T \vec{h_{k-1}} \qquad (1)$$

$$\vec{h_k} = A\vec{a_{k-1}} \qquad (2)$$

上記の定義では収束が不安定となることがある。このため（3）のような安定的なモデルが実用上は使用される。この式においてξは単位ベクトルとAuthorityベクトルのブレンド比率を示し0.95が慣例的に使用される。（3）の定義はPagerankと数学的に同等となる。

$$\vec{a_k} = \xi A^T A\vec{a_{k-1}} + (1-\xi)\vec{e} \qquad (3)$$

次にHISTと企業価値評価の関係文責を行う。分析は以下の線形回帰モデル(4)を用いる。ここで、αは定数、βは回帰係数、ΔHc,IはC社のHITS権威（C社の全特許をまとめた特許ポートフォリオスコア）のi-1年からi年までの変化、i+LagからI+Lag-1年までのC社株価の変化である。LagはC社の開発サイクルを反映したタイムラグであり、0から6までの値である。

$$\Delta S_{c,i+Lag} = \alpha + \beta \log(|\Delta H_{c,i}|) \qquad (4)$$

評価結果を以下の表に示す。**表1**は、東京証券取引所の業種別分類における各グループの割合を示したものである。また、**表2**は、東京証券取引所の定める企業規模に占める各グループの割合を示している。**表1**のCategory of

business typeは業種、Total number of companyは企業数である。**表１**および２におけるＡは0.7以上の相関を持つ比率、Ｂは0.5-0.7の相関を持つ比率、Cは0.5未満の相関を持つ比率である。TOPIX30は、流動性が高く、時価総額の大きい30銘柄で構成されている。TOPIX30に次いで、流動性が高く、時価総額の大きい70銘柄で構成されるのがTOPIX70である。TOPIX Core30とTOPIX Large70の構成銘柄をTOPIX 100とする。TOPIX 400は、TOPIX 500の構成銘柄のうち、TOPIX 100を除いた残りの銘柄で構成される。TOPIX 500の構成銘柄のうち、小型株はTOPIX Smallで指標化されている。

表１．業種ごとの特許引用のHITSスコアとの相関

Category of business type	Total number of company	A	B	C
Pharmaceutical	13	0.10	0	0.90
Electoric power and gas	8	0.24	0.63	0.12
Food	25	0.55	0.14	0.32
Services	2	0.50	0	0.50
Land transportation	5	1.00	0	0
Pulp and paper	5	0.80	0	0.20
Glass	15	0.50	0.17	0.33
Fiber	13	1.00	0	0
Steel	20	0.89	0.06	0.06
Precision Instrument	8	0.50	0.12	0.38
Nonferrous material	11	0.45	0.09	0.45
Automobile	25	0.57	0.24	0.19
Machinery	40	0.72	0.08	0.21
Chemistry	54	0.45	0.22	0.33
Mining	1	1.00	0	0
Construction	39	0.91	0.09	0
IT	3	0	1.00	0
Agriculture and Fishery	2	0.5	0	0.5
Rubber Product	6	0.83	0	0.17
Oil and Coals	4	1.00	0	0
Electronics	62	0.41	0.14	0.44
Metal products	10	0.60	0.2	0.2
Other	18	0.87	0.07	0.07
Warehousing	1	1.00	0	0
Wholesale trade	9	0.86	0	0.14

表２．時価総額ごとの特許引用のHITSスコアとの相関

Scale	A	B	C
TOPIX30	0	0.44	0.56
TOPIX70	0.20	0.35	0.45
TOPIX400	0.50	0.15	0.36
TOPIX Small	0.73	0.10	0.18

結果より概ね企業価値と提案手法の特許スコアの関係性が高いことが示唆されたが、中でも規模が小さい企業群の方が線形回帰モデル（４）に適合する傾向があることが分かった。すなわち、技術力を表す特許スコアは、小規模になればなるほど企業価値に影響を与える可能性がある。また、業種別では、鉄鋼や機械などのBtoB企業では、HITSベースのスコアが株価に相関していることが分かった。

３．引用リンクコミュニティの成長性予測

前章で紹介した引用ネットワークベースの特許重要性評価手法には欠点がある。具体的には静的にスコアを算出しているため衰退産業と成長産業でまったく同じ引用ネットワーク構造になった場合、まったく同じスコアとなることが課題として挙げられる。本章ではそのような欠点を回避するために引用ネットワークの成長性を評価したHentona[13] らが提案した引用リンクコミュニティの成長性予測手法を紹介する。手法の概要を**図２**に示す。まず、特許データベースから引用ネットワークを抽出する。そして、その引用ネットワークから特許コミュニティを検出する。最後に、各コミュニティの成長性を予測するために、LSTM、ARIMA、Hawkes processの３つのモデルの性能を比較した。

大規模なネットワークからコミュニティを検出するためには、ネットワークの各ノードを低次元のベクトルにマッピングすることが重要である。本研究では、Skip-gram[14] でよく利用されるWord2vecのようなNode2vecを用いてこの問題に対処した。Skip-gramは、類似文脈の単語は類似意味を持つ傾向があるという分布仮説[15]、言い換えれば、類似単語は類似単語周辺に現れる傾向が

あることに基づいている。このモデルでは、単語の特徴表現はネガティブサンプリングによるSGDを用いて尤度目標を最適化することにより学習され[16]、計算コストを効果的に削減することができる。Skip-gramをネットワーク情報へ拡張するために、DeepWalk [17] やNode2Vec[18] などの最近の研究では、ネットワークと単語の並びである「文書」の間のアナロジーを用いている。これらの研究は、ネットワークからランダムにノードのシーケンスをサンプリングし、ネットワークをノードとドキュメントの順序付けられたシーケンスに変換する。Node2vecは、DeepWalkの性能を向上させ、近傍ノードをサンプリングする2次ランダムウォーク戦略を設計し、幅優先サンプリング（BFS）と深さ優先サンプリング（DFS）の間を滑らかに補間することができる。BFSでは、近傍ノードは送信元の近傍ノードに限定される。一方、DFSでは、近傍は送信元ノードからの距離が長くなるにつれて順次サンプリングされたノードで構成される。現実のネットワークでは、この２つの属性が混在していることが一般的である。Node2vecは２つのサンプリング戦略を適切に混合するために、ウォークをガイドする２つのパラメータｐとｑを使用する。エッジを通過し、現在ノードに存在するランダムウォークを考える。パラメータｐとｑはBFSとDFSの間を補間するように制御することができ、それによって異なるノードの等価性の概念に対する親和性を反映させることができる。パラメータｐは、探索中のノードをすぐに再訪問する可能性を制御する。これを高い値（＞max（q,1））に設定すると、次の２ステップで既に訪問したノードをサンプリングする可能性が低くなる（ウォークの次のノードが他に隣接していない場合を除く）。この戦略は適度な探索を促し、サンプリングにおける冗長性を回避する。一方、ｐが低い場合（＜min（q,1））、ウォークは一歩後退し、これによりウォークは開始ノードｕの近傍で行われる。パラメータｑは、探索が「内向き」ノードと「外向き」ノードを区別することを可能にする。このようなウォークは、ウォークの開始ノードに関して、小さな局所性内のノードで構成されるサンプルを取得するという意味で、BFSの挙動に近似している。一方、ｑ＜１の場合、ノードｔからより遠いノードを訪問する傾向がある。このような動作は、外側への探索を促進するDFSを反映したものである。

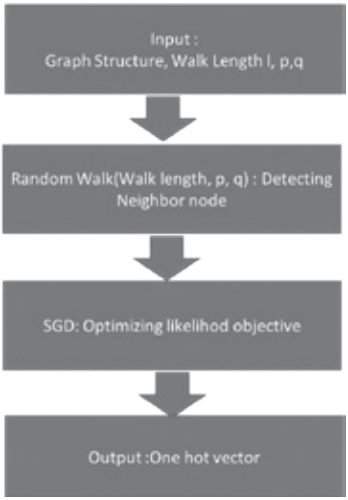

図２．引用リンクコミュニティの成長性予測手法の概要図

Hentonaらの研究では、Node2Vecで生成された特許（ノード）の分散表現ベクトルを用いて階層的クラスタリングを行い引用ネットワーク内の引用集合であるコミュニティを抽出している。そのうえで各コミュニティの成長性をARIMAモデルで予測する。ARIMAモデルは、変数の将来値は、過去値と過去誤差の線形結合で表現したモデルである。ARIMAモデルは伝統的な時系列予測手法であり、様々な分野で応用されている[19)－20)]。ARIMAモデルを選択した理由は、他の有力な時系列予測手法であるLSTMやHawks過程と比較して高い性能を示したためである。**表3**に手法間の結果の比較を示す。

表３．手法間の比較。One Quarter単位の予測。MAPEおよびDirection Accuracyで性能を計測

	Model	MAPE[%]	Direction Accuracy[%]
One Quarter	Hawkes	95.22	43.59
	ARIMA	56.98	82.05
	LSTM	59.63	71.79
One year	Hawkes	88.48	58.97
	ARIMA	71.55	82.05
	LSTM	82.09	64.10

評価結果よりARIMAモデルが最良の予測モデルであることが分かった。これは分や秒単位のデータとことなり時系列方向の数が少ない特許データにおいては時系列方向での大規模データを仮定しているLSTMのような深層学習やHAWKS過程のような確率過程モデルに不向きであり、シンプルなARIMAモデルのほうが高い性能を示しやすいことが背景にある。結論としてNode2VecとARIMAモデルを組み合わせることで技術分野の成長性を解析することが可能であることが示された。

さらにNode2vecの結果を用いて、各クラスタに属する特許の特許文書をランダムに参照し、技術分野を検討した。その結果、共通の技術的なポイントを見出すことができた。パチンコ関連の特許分類では、39のクラスタが存在する。これらのクラスタは、適切にグループ化されている。例えば、あるクラスタは、遊技機およびその周辺技術の不正行為防止に関する特許の集まりである。このクラスタの特許の１つは、スロットマシンの不正改造を困難にする基板ボックス（特願2004-269539）である。この特許は、この基板ボックスを搭載したスロットマシンおよびその技術を応用した特許を引用している。この他、このクラスタの特許として、パチンコ遊技機の検査を容易にする構造、パチンコ遊技機の不正検出機構がある。また、別の例として、遊技機の演出技術に関連する特許の集合体であるクラスタがある。核となる特許は、スロットディスプレイのディープＬ（特願2004-357878）である。この特許は、この方法を実現するための技術特許に引用されている。このほか、クラスタには、スロットマシンの部品構成やパチンコ演出プログラムなどがある。以上の検討により、Node2vecが特許引用ネットワークのグラフクラスタリングに適していることが確認された。

４．まとめ

本論文では引用ネットワークのリンク構造を評価することが特許の価値評価において重要なことを示唆したNonakaらの研究と引用リンクコミュニティの成長性を評価するHentonaらの研究を紹介した。今後、筆者らの研究グループでは両手法を組み合わせた手法の開発を進めていきたいと考えている。

《参考文献》

1）KortumS. , LernerJ. , Stronger Protection or Technological Revolution: What is Behind the Recent Surge in Patenting, National Bureau of Economic Research Working Paper 6204, Cambridge, MA, 1997.

2）A.K. Chakrabarti, I. Dror, Technology transfers and knowledge interactions among defense firms in the USA: an analysis of patent citations, International Journal of Technology Management 9（5）, pp757–770 , 1994.

3）C.L. Tsai, Technology Analysis for Front-end Industry Using Patent Roadmap and Technology Roadmap: A Case Study Based on CNT-FED, Feng Chia University, Taichung, Taiwan, R.O.C, 2006.

4）Kleinknecht A, Van Montfort K, Brouwer E. The non-trivial choice between innovation indicators. Economics of Innovation and New Technology, 11(2), pp.109-121, 2002.

5）Archibugi D, Pianta M. Measuring technological change through patents and innovation surveys. Technovation 16(9), pp.451-468, 1996.

6）Uchida, H. Mano, A. and Yukawa, T.: “Patent Map Generation Using Concept-based Vector Space Model,” Proceedings of the Fourth NTCIR Workshop（2004）

7）Ishikawa, D. Ishizuka, H. Uda, N. and Fujiwara, Y.: “Extraction and Integration of Causal Relationships in Patent Documents: Summary and a Subsequent Activity,” Journal of Japan Society of Information and Knowledge, Vol. 15, No. 3, pp. 98–106(2005)

8）Hirofumi Nonaka, Akio Kobayashi, Hiroki Sakaji, Yusuke Suzuki, Hiroyuki Sakai, Shigeru Masuyama, Extraction of the Effect and the Technology Terms from a Patent Document, Journal of Japan Industrial Management Associastion, 63, pp.105 - 111, 2012.

9）Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. "Market value and patent citations." RAND Journal of economics（2005）: 16-38.

10）Jon Kleinberg,“Authoritative sources in a hyperlinked environment”, Journal of the ACM,No.46(5), pp.604-632. 1999.

11）Lukach, Ruslan, and Maryna Lukach. "Ranking USPTO patent documents by importance using random surfer method（PageRank）." (2007).

12）Nonaka, H., Kubo, D., Kimura, T. H., Ota, T., & Masuyama, S.（2014, June）. Correlation analysis between financial data and patent score based on HITS algorithm. In 2014 IEEE International Technology Management Conference（pp.1-4）. IEEE.

13) Hentona, A., Nonaka, H., Nakai, K., Sakumoto, T., Kataoka, S., Alemán Carreón, E. C., ... & Hirota, M. (2018, September). Community detection and growth potential prediction from patent citation networks. In Proceedings of the 10th International Conference on Management of Digital EcoSystems (pp. 204-211).

14) Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp.3111-3119).

15) Z. S. Harris., Distributional Structure, Word 10(23):146–162, 1954.

16) T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

17) Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

18) Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855-864, 2016.

19) Bakar, Nashirah Abu, and Sofian Rosbi. "Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction." International Journal of Advanced Engineering Research and Science 4.11 (2017).

20) Hernández, Nathalie, et al. "Arima as a forecasting tool for water quality time series measured with UV-Vis spectrometers in a constructed wetland." Tecnología y Ciencias del Agua 8.5 (2017): 127-139.