社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－139 (2)
2006－ＦＩ－ 83 (2)
2006／5／30

引用情報に基づく特許文献の重要度算出方式の検討

佐藤 祐介　　岩山 真
{y-sato, iwayama}@crl.hitachi.co.jp
日立製作所 中央研究所
〒185-8601 東京都国分寺市東恋ヶ窪 1-280

本研究では、特許文献間の引用関係を解析することで重要特許を発見する方法について検討した。他の特許からの引用が多いほど重要特許である可能性が高いと仮定し、引用関係に基づいて特許の重要度を計算する。本研究では、特許の質的評価に有効となる情報としてどういったものが有効であるかを検討した。その結果、「出願人引用の方が審査官引用よりも特許の重要度計算に適している」、「自社引用の方が他社引用よりも特許の評価に有効である」といった知見を得ることができた。実験により多くの重要特許サンプルがそれ以外の特許よりも重要度が高くなることを確認した。最大で上位50位までに約65%のサンプルが現れる結果を得た。

A Study of Patent Document Score Based on Citation Analysis

Yusuke Sato　Makoto Iwayama

Hitachi, Ltd., Central Research Laboratory

1-280, Higashikoigakubo, Kokubunji, Tokyo, 185-8601 Japan

This paper describes experiments for defining the significance score of patent by using patent citations. We assume that the more highly a patent is cited, the higher significance score the patent has. Based on this assumption, we investigated what kind of feature about patent citation is effective for detecting important patent. Through the experiments using Japanese patents, we found that the citation by applicant is more effective factor than the citation by examiner. We also found that the self citation is an important factor. By incorporating these findings to the significance score, about 65% samples of important patents appeared within the top 50 ranking.

1. はじめに

近年日本では、アジア各国・地域の企業の急成長を背景として、国内産業の国際的競争力を高めるために知的財産を適切に保護・活用していく方策が国家的に推し進められている。こういった中、大学等の学術機関や企業は知的財産の調査・管理に大きな力を注いでおり、競合相手の特許出願状況を的確に把握するための分析ツールの重要性が増している。特に、重要特許を特定するための機能は、公知例調査の検索作業に費やす時間の短縮につながるため、応用の可能性が高い機能の1つである。

技術文献の重要度には被引用数を用いる方法が一般的である。米国では、米国特許庁が早くから特許の引用情報の整備を行ってきたため、引用情報を利用した解析が盛んに行われてきた[2][3]。一方日本では、平成14年9月の先行技術文献情報開示制度により初めて引用文献の明細書記載が義務化されたということもあり、引用情報を用いた特許の重要度の評価研究は前例が少ない[4][5]。

本研究では、日本国特許を対象に引用情報を利用した特許の重要度を測るための尺度について検討した。従来の被引用数に基づく評価方法に則りつつも、特許に固有の情報の活用可能性とその方法について検討を行った。

2. 特許引用を用いた重要度算出方法決定のアプローチ

2.1. 特許引用分析に関する従来研究

1980年以降、特許の引用に関するデータベースが米国特許において整備されたのをきっかけに、米国特許の引用を用いたさまざまな分析が盛んとなった。例えば、Albertらは特許の被引用数と技術者による革新度の評価との相関の高さを示した[2]。Carpenterらは著名な賞を受賞した技術に関する特許はその他の特許よりも被引用数が多いことを実証している[3]。

一方、日本国特許に関しては、内藤による888万件の特許をデータベース化し引用関係のネットワークの性質を調査した研究[5]や、小川らによる審査官と出願人による引用や自社／他社引用の特許重要度との関係を検証した研究[4]が報告されている。しかし、これまでの研究はある特定の技術における重要特許と被引用数の関係を検証したものであり、より詳細な関係の傾向を捉えるためには多くの技術分野の重要特許を用いた検証が必要と考える。

本章では次節から、本研究において検証した特許の引用に固有の情報の活用方法とそれらを特許重要度へと応用した過程を述べる。

2.2. 特許の引用に固有の情報

本節では、本研究で検討した特許に固有の情報について説明する。

i. 引用種

特許文献の引用には、審査請求された特許が審査官により拒絶された場合に、その理由として通知される「審査官引用特許」と、出願人が特許明細書中で引用している「出願人引用特許」があり、この違いによる重要度の変化を検証した。

ii. 会社別引用

出願人引用は、引用先と元の特許の出願人（会社）が同じである自社引用と、それらが異なる他社引用の2種類に分けられる。特許の執筆者である発明者とは別に、出願人という属性があることで“自社／（同業）他社からの引用”といった分類ができることは特許の引用に固有の情報である。一般的に自社引用は他社引用に比べて多く、この引用の偏りが重要特許の評価に悪影響を及ぼしてしまうことが予想される。したがって、自社引用のみの重要度、

他社引用のみの重要度、それら両方を含んだ重要度のうち、どの重要度が適しているかを比較した。なお、本研究では個人や大学の出願人も会社とみなしている。また、グループ企業内会社間の引用は他社引用として扱っている。

iii. 引用単位

引用されている特許文献の数を被引用数とした「特許数」と、引用されている出願人（会社）の数を被引用数とした「会社数」の違いによる特許重要度の変化を検証した。会社数は、より多くの会社から注目を集める特許ほど重要特許の可能性が高いと仮定して採用した。例えば、ある特許の被引用特許がA（X社）、B（Y社）、C（Y社）、D（Z社）、E（Z社）であった場合は、特許数が5件、会社数が3件となる。

以上をまとめると、本研究で用いた特許固有の情報は表 1のようになる。これらを重要度計算の際用いる引用の条件と位置付け、どの条件の組み合わせが最も適しているのかを検証する。組み合わせとは例えば、引用種に「出願人引用」を用い、「自社＋他社引用」による、「引用単位を会社数」とした条件といったものである。

表 1：特許重要度に用いた特許の引用に固有の情報

#	特許の引用に固有の情報	特許重要度への適用方法
i.	引用種	出願人引用
		審査官引用
ii.	会社別引用	自社＋他社
		自社のみ
		他社のみ
iii.	引用単位	特許数
		会社数

2.3. 重要度の算出式

前節で説明した特許の引用に固有の情報を用いて、どのような算出式が重要度として有効なのかを検討した。

① 被引用数

特許集合全体から抽出する被引用の数。

② 平均被引用数

被引用数を公開年から現在（本研究では2005年とする）までの経過年数で割った値。出願の古さに被引用数が比例してしまうことを防ぐための算出式。Nをその特許の全被引用数とし、xをその特許の公開年とすると、以下の式により求めた値E_2：

$$E_2 = N/(2005 - x).$$

③ エントロピー

2つの特許の被引用数が同じであれば、ある年に被引用が集中している特許よりも、公開から現在に至るまで均等に引用されている特許の方が重要性が高いのではないかという仮説に基づいて採用した。

n(x)をx年における被引用数、Nをその特許の全被引用数とし、以下の式により求めた値E_3：

$$E_3 = \sum_x - p(x)\log p(x) \quad (\because p(x) = n(x)/N)$$

④ HITS

引用／被引用先の特許の重要度も考慮した計算方法として HITS（Hyperlink Induced Topic Search）[1]を採用した。WWW のハイパーリンク関係を解析する手法で、「多くの良質なリンクを持つページ（“Hub”）からリンクされているページ（“Authority”）は、良質なページである」という再帰的な関係を前提として、良質な引用をもつページの評価値である Hub と良質な被引用をもつページの評価値である Authority の 2 種の重要度を計算する。本研究では、被引用に基づく評価値である Authority 値が高い特許を重要であるとみなした。

2.4. 条件と算出式の組み合わせ

2.2 節の特許固有の情報を条件とし、2.3 節の算出式との組み合わせにより最終的な重要度を計算する。例えば、「審査官引用、他社引用のみ、引用単位を会社数とした条件の下、平均被引用数を用いて計算した重要度」といった組み合わせである。ただし、エントロピーと HITS には「引用単位」の条件を適用していない。複数ある同一会社の引用をどの引用で代表するかによって最終的に得られる重要度が変わってしまうのを避けるためである。例えば、エントロピーの場合は同一会社の引用が複数年にある場合、どの年で代表するかで値が変わってしまうからである。

最終的に、全部で 32 種類の算出方法となる（全組み合わせは後に示す表 3 を参照されたい）。本研究では、実験によりこれらの中から最も適した特許重要度算出方法を探した。

2.5. 実験の手順

実験には 1993 年から 2002 年までの公開公報データを用いた。重要特許のサンプルとして(社)発明協会が発表する各種発明賞を受賞した特許 181 ケースのうち、公開年が 1993 年から 2002 年である 48 ケースを用いた。この受賞特許のことを「正解特許」と呼ぶこととする。

これらのデータを用いて、32 種類の特許重要度算出方法のうち、より多くの正解特許をより上位に押し上げることができる算出方法を探す。本実験は「評価対象集合の抽出」と「正解特許順位の集計」の 2 つのステップに分けられる。

◆ 評価対象集合の抽出

本ステップでは、正解特許と同じ技術テーマに属する特許を集めた評価対象集合の抽出を行う。正解特許の第一請求項から概念検索を行い、検索結果上位 10000 件を集め、その上位から順に、引用もしくは被引用をもつ 300 件を抽出して評価対象集合とした。正解特許が属する個々の分野に限定して評価対象集合を作成したのは、特許には数多くの技術分野があり、その各分野において重要特許が存在することを仮定しているからである。特許集合全体で一律に評価してしまうと、異分野の重要特許を比較するという困難な問題が生じてしまう。

◆ 正解特許順位の集計

まず、評価対象集合中の特許それぞれの重要度を本研究で採用した算出方法を用いて計算する。そして、重要度に従って順位付けを行い、正解特許がランキングの何位であるかを調べる。ここで、実際にランキングを行うと、下位において被引用数が同じなために同位となる特許が大量に出てしまい、精度の差がほとんど見えない重要度が多く見られた。このため、被引用数に関する重要度の場合はエントロピー値により、それ以外の重要度は被引用数により同順位の特許の再ランキングを行った。

次に、これまでの操作を正解特許 48 ケースそれぞれについて行い、正解特許の順位の累積度数を求める（図 1)。そして、この累積度数の計算を 32 種類の重要度算出方法について行い、どの方法がより多くの正解特許をより上位にランキングしているのかを検証する。

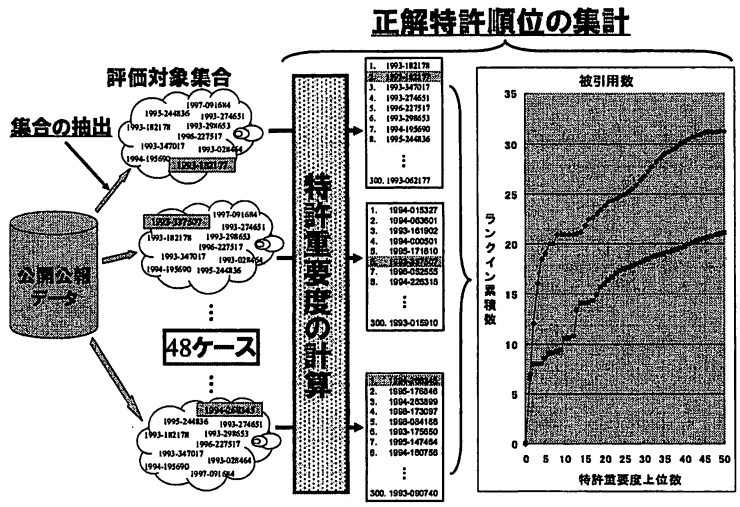

図 1：正解特許の順位集計の流れ

3. 実験結果及び考察

本節では、これまでに述べた特許固有の条件と算出式を用いて行った重要度の実験の結果について示す。初めに条件と算出式の全組み合わせによる結果を俯瞰した後、個々の組み合わせの有効性について述べる。

3.1. 全体結果

まず初めに、本研究で用いた 10 年分の特許に関する統計値を表 2 に示す。なお、本研究での被引用数とは、引用されている公開公報の数である。海外も含めた特許文献の引用は用いていない。また、出願人引用はパターンマッチにより明細書中から抽出した。

表 2：本研究で用いた 10 年分の特許の引用データ

	出願人引用	審査官引用
引用がある公報	871,111 件	868,764 件
被引用がある公報	765,503 件	1,510,001 件
総公報数	3,496,252 件	

出願人引用よりも審査官引用の被引用数の方が多いことがわかる。

表 3 に全 32 種類の特許重要度の実験結果を示す。この表は、それぞれの重要度のランキングにおける、「上位 50 位までの正解特許順位累積数」、「50 位における再現率」、「1 位から 50 位までの再現率の平均」を示している。なお、再現率の平均とは、1 位から 50 位のそれぞれの点における再現率を平均したものであり、以下の式により求めている。

$$\frac{1}{50}\sum_{i=1}^{50} r_i \quad (r_i : i \text{ 位における再現率})$$

上位50位による結果を示したのは、公知例を可能な限り漏れなく探すという特許検索の性質上、ユーザは最低でも50件程度の検索結果を必ず閲覧するということを想定しているからである。よって、重要度順で上位 50 位以内の再現率が重要であると考えた。しかし、表 3 を見てもわかるとおり、50 位点での再現率だけでは評価が難しい場合が多い。よって、50 位での再現率が同じ場合は、再現率の平均を比較することで、より多くの正解特許をより上位にランキングした重要度算出方法を見つけ出す。

表 3：32種の特許重要度算出方法の累積数・再現率・再現率の平均の比較結果

			引用種					
			出願人引用			審査官引用		
算出式	会社別引用	引用単位	累積数	再現率	平均	累積数	再現率	平均
被引用数	自社＋他社	特許数	31.23	0.651	0.520	21.09	0.439	0.331
		会社数	31.23	0.651	0.499	21.80	0.454	0.297
	自社のみ		26.36	0.549	0.471	18.80	0.392	0.306
	他社のみ	特許数	22.00	0.458	0.346	15.70	0.327	0.253
		会社数	22.00	0.458	0.354	17.20	0.358	0.259
平均被引用数	自社＋他社	特許数	31.00	0.646	0.508	19.00	0.396	0.302
		会社数	26.00	0.542	0.379	18.61	0.388	0.240
	自社のみ		26.36	0.549	0.459	16.96	0.353	0.290
	他社のみ	特許数	20.00	0.417	0.318	15.00	0.313	0.221
		会社数	20.00	0.417	0.300	13.00	0.271	0.207
エントロピー	自社＋他社		31.23	0.651	0.515	21.23	0.442	0.306
	自社のみ		26.36	0.549	0.468	18.80	0.392	0.305
	他社のみ		31.23	0.651	0.471	18.41	0.384	0.243
HITS	自社＋他社		31.10	0.648	0.542	19.58	0.408	0.333
	自社のみ		26.36	0.549	0.471	18.39	0.383	0.314
	他社のみ		22.00	0.458	0.343	15.00	0.313	0.259

各重要度とも差はわずかであるが、再現率の平均値から「自社＋他社引用を用いた HITS による重要度算出方法が良い」とみなすことができる。

3.2. 特許の引用に固有の情報の比較

■ 引用種

図 2 は条件を自社＋他社引用、引用単位を特許数に固定し、算出式を被引用数として、引用種の違いによるランクインの累積ケース数を比較したグラフである。横軸が重要度順に特許を並べた場合の順位、縦軸がその順位までにランクインした正解特許の累積ケース数である。つまりは「どれだけ多くの正解特許がより上位にランクインしたのか」を示したグラフとなる。したがって、上に位置するプロットほど良い算出方法であることを表す。

出願人引用の方が審査官引用よりも多くの正解特許を上位に位置づける結果となった。審査官と出願人では引用の観点が違っていることが理由として考えられる。審査官が引用に用いる公報は、技術的に優れた公報というよりは、権利を無効化できる公報である場合が多いが、出願人は技術的に優れた公報を引用する傾向が強いため、結果的に審査官引用よりも良いランキングとなったと考えられる。

■ 会社別引用

図 3 は条件を出願人引用に固定し、算出式を被引用数として、ランクインの累積ケース数

を比較したグラフである。縦軸、横軸は図 2 と同じである。自社／他社引用のみでの算出方法では正解特許のランキングが低下する結果となった。特に、他社引用のみによる結果はランキングの低下が著しい。これは、自社で重要と認めた特許ほど防衛のために周辺特許を数多く出願する傾向があり、自社引用が重要特許を評価する上で有効なファクターであることを示していると考えられる。表 4 は正解特許とそれ以外の特許における全被引用数に占める自社引用の割合の平均を示したものである。特に出願人引用では、正解特許における全被引用数の約半数が自社引用であるのに対して、その他の特許では 4 分の 1 程度であり、自社引用の割合の多さが特許重要度を決定する有効な情報の 1 つであることがわかる。

表 4：自社引用の割合の平均

出願人引用	正解特許のみ	56.0%
	評価対象集合全体	27.4%
	評価対象集合全体（正解特許除く）	26.9%
審査官引用	正解特許のみ	39.6%
	評価対象集合全体	19.7%
	評価対象集合全体（正解特許除く）	19.5%

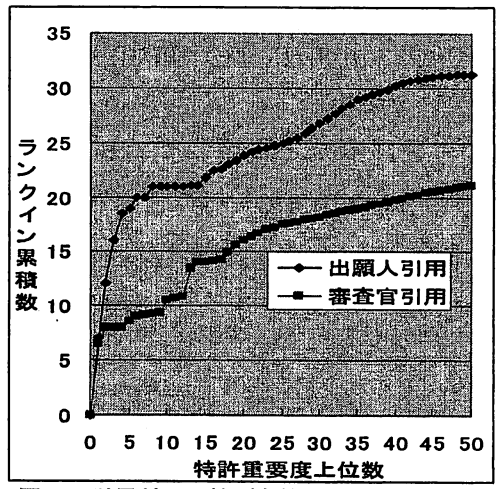

図 2：引用種の比較（自社＋他社、引用単位を特許数とし、被引用数に限定した場合）

■ 引用単位

「自社＋他社」条件との組み合わせでは引用単位を“特許数”とする方がわずかに良い結果となった。しかし、「他社のみ」条件との組み合わせにおいては差が現れなかった（図 3）。これについても上記で述べた「会社別引用」の比較と同じことが言える。つまりは、自社引用数がのべ数から 1 に丸め込まれたために、“多くの周辺特許により重要特許を防衛している”という情報が落ちてしまい、ランキングが低下したものと考えられる。

3.3. 算出式の比較

図 4 は、条件を出願人引用、自社＋他社引用、引用単位を特許数として 4 種の算出式を比較したグラフである。わずかな差で HITS、被引用数、平均被引用数、エントロピーの順に良い結果となっているが、これらの算出式に関する有意差は見られなかった。

各特許におけるランキングを詳しく調べると、被引用数では上位にあるものがエントロピーでは大きく順位を落とす場合や、またその逆のパターンになる場合が見られた。エントロピーが高い公報ほど多年にわたって引用されている傾向が強く、当初意図した結果が出ていることが確認できた。これについては、「被引用数」と「エントロピー」の 2 つを組み合わせることで、“被引用数”と“年数による広がり”の 2 つの意味を持った指標が可能かどうかを検証する必要がある。また、エントロピーについては正規化も行ったが、順位に大きな変化は見られなかった。平均被引用数では特に大きな特徴は現れなかった。

4. おわりに

特許文献間の引用関係を解析することで重要特許を探索する方法について検討した。特許に固有の情報の活用を考慮することで、重要特許のサンプルをより上位に位置させる文献重要度の評価方法を検討した。その結果、特許の引用に固有の情報では

- 審査官引用よりも出願人引用の方が重要特許の評価に適している
- 自社引用が重要特許を評価する上で重要である

という知見が得られた。また、重要度算出方法の中では、自社＋他社引用を用い、引用元／先の文献の質を考慮するHITSによる順位が最も良く、上位50位以内に48件中31件（約65%）が現れる結果となった。

今後は、例えば海外出願の有無といったより詳細な特許書誌情報の利用による重要度の精度改善や、被引用数による評価が困難な公開年の新しい特許の評価方法について検討する必要がある。

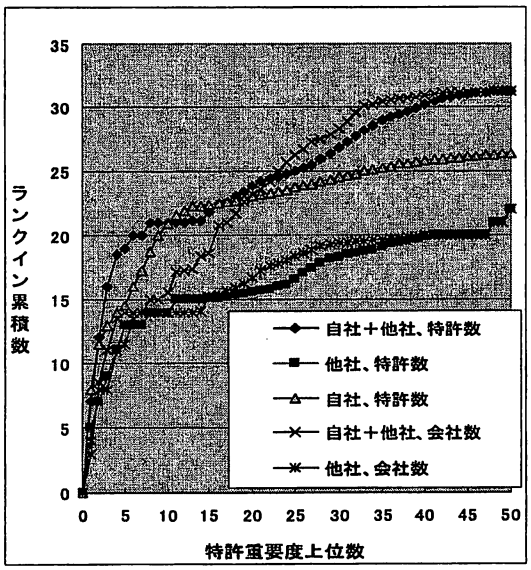

図 3：出願人引用、被引用数に限定した場合の結果

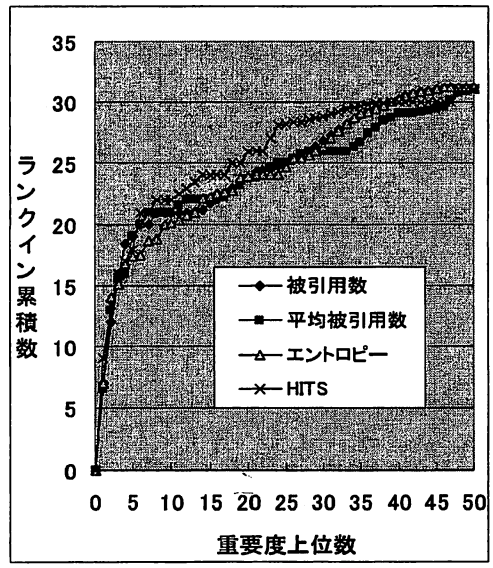

図 4：算出式の比較（出願人引用、自社＋他社、引用単位を特許数とした場合）

参考文献

[1] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *J. ACM* 46(5), pp.604-632, 1999.

[2] M. B. Albert, D. Avery, F. Narin & P. McAllister, Direct validation of citation counts as indicators of industrially important patents, *Research Policy* 20 pp.251-259, 1991.

[3] M. C. Carpenter, F. Narin & P. Woolf, Citation Rated to Technologically Important patents, *World Patent Information* 4 160-3, 1981.

[4] 小川知也、渡部　勇、“引用情報に基づく基本特許抽出”、情報処理学会研究報告　情報学基礎、2005-FI-078、pp.41-48、2005.

[5] 内藤祐介、“特許引用関係ネットワークにおける技術進化の分析方法”、日本ソフトウェア科学会　ネットワークが創発する知能研究会　第一回ワークショップ　WEIN2005講演論文集、pp.51-54、2005.