様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：１２６０８
研究種目：基盤研究（Ｃ）
研究期間：2010～2012
課題番号：22500848
研究課題名（和文）　科学と技術のリンケージの分析を可能とする情報基盤整備に向けた基礎的研究
研究課題名（英文）　Study on Approach to Identify Science Publications within Non-patent References in Patents
研究代表者
調　麻佐志（SHIRABE　MASASHI）
東京工業大学・大学院理工学研究科・准教授
研究者番号：00273061

研究成果の概要（和文）：
科学と技術の連関の分析に活用される特許に引用された科学論文の DB 構築のため、特許上の引用文献と論文 DB 収録文献を実用レベルの精度で照合する手法を開発した。具体的には、曖昧さを許容した項目毎の照合を行い、それらの照合を組み合わせることで正確さを担保した照合アルゴリズムを作成し、さらに個々のアルゴリズムではカバーできない照合を教師データに対する学習に基づく複数アルゴリズムの組み合わせによって可能とした。

研究成果の概要（英文）：
For building a database of linkages between science publications and non-patent references in patents, I developed an approach (i.e., a program) to identify such linkages at practical level. This program is an optimized combination of algorithms to match non-patent references to records of a publication database (SCI-Expanded) by using several match-keys.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	700,000	210,000	910,000
2011 年度	500,000	150,000	650,000
2012 年度	600,000	180,000	780,000
年度			0
年度			0
総　計	1,800,000	540,000	2,340,000

研究分野：総合領域
科研費の分科・細目：科学教育・教育工学
キーワード：科学技術政策

１．研究開始当初の背景

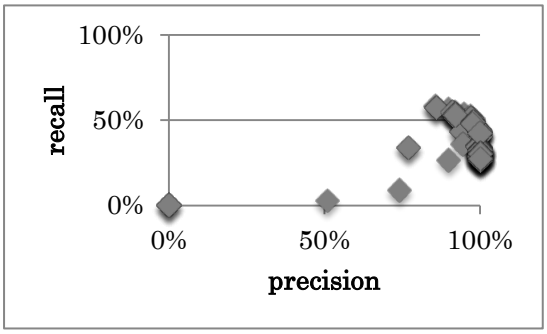
各国で科学技術政策の焦点がイノベーションに移行した中で、科学研究がイノベーションに果たす役割をいかに把握し、解析するかという研究課題は、これまで以上に重要性を増している。この課題を取り扱うために重要なツールないし概念の一つとして特許による学術論文の引用（サイエンスリンケージ）があり、その統計的分析が活発に行われている。しかし、サイエンスリンケージのデータベースには、極めて高価かつ精度が不明な商用データベースしかなく、そのことがこの領域における研究の阻害要因となっていた。

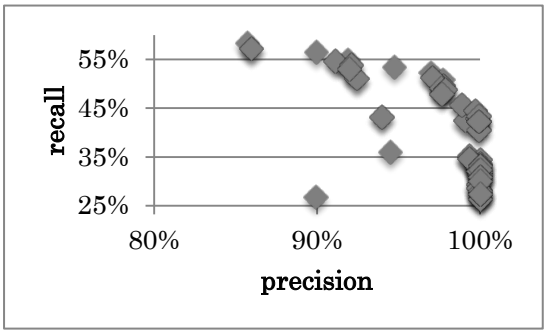
サイエンスリンケージの定量的分析については、既に様々な先行研究があるが、そのなかでは米国の Narin らの一連の研究が先駆

的である。Narin らの研究は、米国特許データベース内の NPR（非特許引用文献）データを分析対象としており、科学計量学の分野でよく用いられる文献データベースである SCI と NPR データの照合を行うことにより、特許に引用された科学論文の国別シェアなどのデータを算出している。この分析により、1980 年代後半以降のサイエンスリンケージの急増が示されたことは、イノベーションに果たす科学の役割に関する科学技術政策上の議論に大きな影響を及ぼした。しかし、Narin らの論文では、SCI との照合の具体的な手法が述べられていないだけでなく、マッチングに成功した論文の数や割合についての具体的な情報を示していない。そのため、この分析結果が、どの程度の精度を持つものかは明らかでない。
サイエンスリンケージの分析を行うためには、引用された科学論文の書誌情報に基づく集計が必要であるが、特許における NPR データは、書誌情報として充分でなく、また質的にも問題があることが知られている。すなわち、情報は標準化されておらず精度も決して高くない。このようなデータの質の問題は、当然、分析の障害となる。たとえば、特にサイエンスリンケージの分析における基本指標である「特許による科学論文の被引用回数」を集計するためには、それぞれの NPR データが論文として一意的に同定されることが必要であるにも関わらず、それを行うことは非常に困難である。
この種の照合は、すなわち、大規模かつ標準化されていない NPR のデータと同じく大規模な文献データの論文レベルでの突合作業を意味するが、現状ではコスト面で障害を抱える作業となっている。確かに技術分野を限って少数の NPR を抽出して人手に頼りこれら作業を完遂することはさほど困難ではなく、既存のサイエンスリンケージに関連する研究もそのような研究がほとんどである。しかし、今後の科学技術イノベーション政策研究の進展に必要な網羅的な分析を実施しようとすれば、一気にハードルが上がる。このような照合を自動的に高い精度で実現する手法の開発が待たれていた。

２．研究の目的
本研究においては、科学技術政策研究において主要なトピックスの一つである科学と技術のリンケージを解析する際に活用される米国特許に引用された科学論文のデータベースの構築に向けて、米国特許に記載された引用文献と学術論文データベースに収録された文献との間の照合自動化を実現すること、すなわちリンケージを同定する実用的なプログラムの開発を目的とした。

３．研究の方法
リンケージを同定するために、まず米国特許に記載された引用文献情報 20,000 件を（擬似乱数による）「ランダム」サンプリングによって抽出し、それに対して手作業で学術論文データベースとの照合を行い、その結果の半分を教師データ、残りの半分を評価データとした。この引用文献情報に対して、曖昧さをある程度許容する形で match-key 毎の照合を行い、それら項目（筆頭著者姓、巻、最初のページ、発刊年、論文タイトル、掲載誌名など）の照合結果を組み合わせることで正確さを担保した個別アルゴリズムを作成した。これら個別のアルゴリズムの精度は下図のように分布している（下は主要箇所を拡大したもの）。





ついで、これら個々のアルゴリズムではカバーできないリンケージの同定を教師データに対する学習に基づいた複数の個別アルゴリズムの組み合わせにより実現した。さらに、この最適化された組み合わせのパフォーマンスを、評価データを用いて評価した。最適化した際の教師データに対するパフォーマンスは次ページ表左の通りである。また、独立した評価データによる評価結果は、次ページ表右の通りである。なお、表内の組み合わせについて、各々、A は precision（正確さ）を追求した組み合わせ、B は precision と recall（カバレッジ）のバランスをとったもの、そして C は recall を重視したものである。
表に示された照合の精度は、少なくとも先行研究を大きくアウトパフォームしたレベルにあり、十分に実用レベルに到達していると判断できる。

最適化の結果			評価データによる評価結果	
組み合わせ	正確さ	カバレッジ	正確さ	カバレッジ
A	100%	84.8%	99.9%	83.6%
B	99.6%	92.8%	99.5%	91.8%
C	95.2%	95.5%	94.0%	94.5%

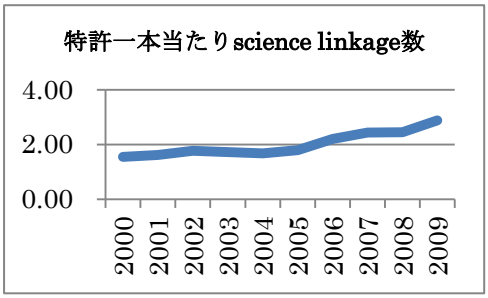
４．研究成果

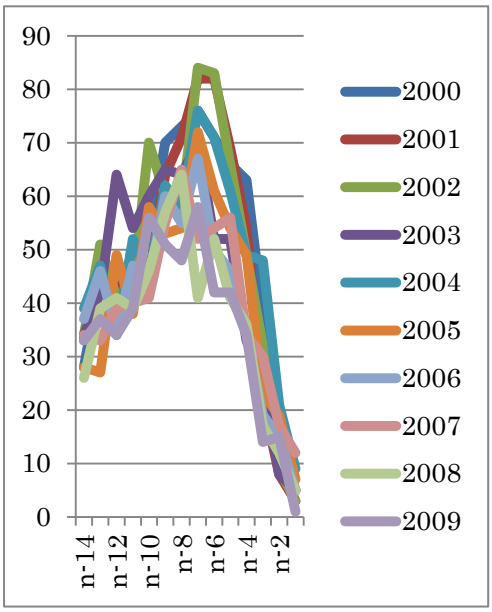
①サイエンスリンケージの基礎統計（推計値）評価・教師用データを作成した結果、米国特許による論文引用に関する基礎的な統計が推計可能となった（下表）。その結果、米国特許一本あたりに記載される非特許引用文献数が急増しているため、非特許引用文献に占める学術論文のシェアこそ低下しているもの、同じく特許一本あたりのサイエンスリンケージは上昇傾向にあることなどが確認された（右図）。

登録年	*2000*	*2001*	*2002*	*2003*	*2004*
特許数	157496	166038	163518	169035	164291
NPR数	466056	519743	549741	585150	557524
うち論文	1042	1032	1040	1012	983
同%	52.1%	51.6%	52.0%	50.6%	49.2%
年	*2005*	*2006*	*2007*	*2008*	*2009*
特許数	143806	173770	157283	157772	167349
NPR数	557780	851232	868929	936926	1139407
うち論文	932	922	895	844	835
同%	46.6%	46.1%	44.8%	42.2%	41.8%

さらに特許登録年と当該特許に引用された学術論文の出版年の間のラグのピークが６～７年であることを明らかにした。

②特許と文献の実用レベルの照合方法の開発





米国特許に記載された非特許引用文献と学術論文データベースに収録された文献との間の照合を機械的に行う手法を研究し、それを実装したプログラムを開発した。現段階において、マッチングの適切さは、precision（正確さ）で 99.5%、recall（カバレッジ）で 91.8%と評価されており、先行研究の成果を大きく凌駕しており、実用的なマッチング手法を開発するという当初の研究成果目標を達成した。

５．主な発表論文等

〔雑誌論文〕（計１件）

① 調麻佐志（印刷中），科学計量学と評価、科学技術社会論研究、第 10 号（査読付き）.

〔学会発表〕（計４件）

① Masashi SHIRABE（accepted、ポスターセッション），“Building an “available” dataset of scientific citations in patents for comprehensive analyses and indicators of S&T interactions”, 18th International Conference on Science and Technology Indicators（2013 年 9 月 4～6 日、ベルリン）.

② Masashi SHIRABE（accepted、査読付き full paper），“Approach to Identify

SCI Covered Publications within Non-patent References in Patents”, ISSI 2013 (2013年7月15〜18日、ウィーン大学).

③ 調麻佐志 (2012), 米国特許が引用する学術論文の計量書誌学的分析, 研究技術計画学会第27回年次学術大会 (2012年10月27〜28日、一橋大学).

④ Masashi Shirabe (2010), “Analysis of patent citations of scientific articles and its implication from STS perspective”, 国際科学技術論学会・科学技術社会論学会合同会議 (2010年8月27日、 東京大学駒場キャンパス).

６．研究組織

(1)研究代表者

調　麻佐志 (SHIRABE　MASASHI)

東京工業大学・大学院理工学研究科・准教授

研究者番号：00273061

(2)研究分担者

該当なし。

(3)連携研究者

該当なし。