

Title	特許文献を利用した研究者逆引きデータベースシステムの構築
Author(s)	橋本, 泰一; 乾, 孝司; 内海, 和夫; 石川, 正道
Citation	年次学術大会講演要旨集, 24: 271-274
Issue Date	2009-10-24
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/8626
Rights	本著作物は研究・技術計画学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Science Policy and Research Management.
Description	一般講演要旨

特許文献を利用した研究者逆引きデータベースシステムの構築

○橋本 泰一（東京工業大学 統合研究院）
 乾 孝司（筑波大学 大学院システム情報工学研究科）
 内海 和夫（東京工業大学 統合研究院）
 石川 正道（東京工業大学 大学院総合理工学研究科）

1. はじめに

学際的研究や産学官連携による共同研究の促進により、複数の分野の研究者が共同で研究を行う機会の増加にともない、網羅的な研究者の探索が必要とされている。そのため、研究者のプロフィールや研究業績などの研究者情報が網羅的に収集され、研究者探索のサービスの提供が行われるようになった。日本国内のサービスとしては、科学技術振興機構が提供する研究開発支援総合ディレクトリ ReaD¹や JST 文献検索サービス JDreamII²、国立情報学研究所が提供する NII 論文情報ナビゲータ CiNii³や科学研究費補助金データベース KAKEN⁴、大学をはじめとする各研究機関が提供する機関レポジトリ⁵がある。海外においては、トムソン社が提供する Web of Science をはじめとする学術文献情報データベース⁶、エルゼビア社の SCOPUS⁷などが挙げられる。

本論文では、キーワードを入力し、そのキーワードに関連した研究者を出力するデータベースを研究者逆引きデータベースと呼ぶ。前に挙げた各サービスでは、キーワードにより研究者の研究活動に関連したコンテンツを検索することにより、キーワードに関連した研究者を探索でき、研究者逆引きデータベースとしての機能は果たしている。しかし、二つの問題点がある。一つは、これらのデータベースのコンテンツは主に学術論文などの研究者の研究活動の結果であるため、検索に利用できるキーワードは学術に関連したキーワードに限られる。もう一つは、研究者が今までに研究していないが、十分取り組むことが可能な新たな研究課題に関するキーワードでは研究者を発見することができない。

我々は、この二つの問題を解決するために新しい研究者逆引きデータベースを提案する。我々のデータベースは、特許と研究者をキーワードにより関連づけ、キーワードにより特許検索結果から関連研究者を出力する。具体的には、事前に、研究者の学術論文から研究者の研究テーマを表す代表的なキーワードを抽出し、それらのキーワードを多く含む特許文献を研究者の研究テーマと関連する特許であると見なす研究者-特許文献関連度を計算しておく。そして、ユーザがキーワードにより特許検索を行う際に、検索結果の上位の特許文献と関連度が高い研究者をランキングし出力する。実際に、2006 年度に東京工業大学に所属していた 125 名の研究者と 2004 年から 2007 年の公開特許公報との関連度を計算し、約 20 万件の特許文献に対して研究者を関連づけたデータベースシステムを構築した。

2. 社会課題解決支援システム RiverStone

我々は、社会課題を発見し、その課題を解決するための技術要素、その技術要素を研究・開発する能力を有する研究者の探索を支援するデータベースシステム（RiverStone）の構築を目指している。RiverStone では、新聞記事、特許、学術論文を収録し、各コンテンツの検索機能を提供している[橋本 08a]。加えて、新聞記事においては、文書検索、文書クラスタリング、文書要約の 3 つ機能により社会課題発見のための新聞記事分析を支援する[橋本 08b, 乾 08]。これまで安全安心に関する社会課題に関する分析[橋本 08c]や医療技術に関する分析[内海 09]を報告してきた。

¹ <http://read.jst.go.jp/>

² <http://pr.jst.go.jp/jdream2/>

³ <http://ci.nii.ac.jp/>

⁴ <http://kaken.nii.ac.jp/>

⁵ <http://www.nii.ac.jp/irp/>

⁶ <http://www.thomsonscientific.jp/>

⁷ <http://www.scopus.com/>

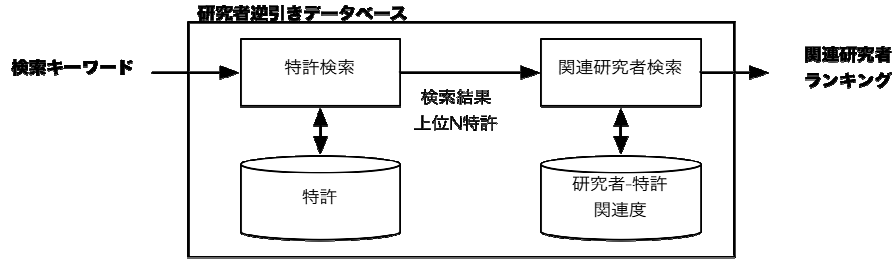


図 1 特許文献-研究者関連度の計算方法

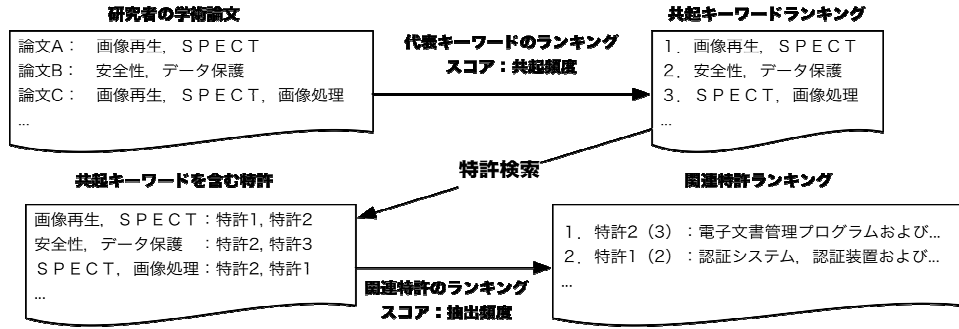


図 2 研究者逆引きデータベースシステムの概要

3. 特許文献を利用した研究者逆引きデータベースシステム

3.1. 特許文献を利用した研究者逆引きデータベースシステムの概要

RiverStone における研究者逆引きデータベースは、入力されたキーワードにより特許検索を行い、その検索結果上位の特許文献と関連の深い研究を行った研究者を推定しランキング形式で出力する(図 1)。研究者のランキングに用いるスコアは、研究者と検索結果上位 n 件の特許との関連度の総和を用いる。

$$\text{Score}(r) = \sum_{i=1}^n \text{Rel}(r, p_i) \quad (1)$$

研究者 r のスコアは、上記の式(1)により計算される。 $\text{Rel}(r, p_i)$ は、研究者 r と検索結果 i 番目の特許文献 p_i との関連度を表し、次節に定義する。

3.2. 研究者-特許文献の関連度

研究者と特許文献を結びつけるために、研究者の研究業績の一つである学術論文を利用する。論文の書誌情報より研究者の研究活動を表す代表的なキーワードを抽出し、それらのキーワードを多く含む特許文献を研究者と高い関連がある特許と考える。その計算方法の概要を図 2 に示す。

科学技術振興機構が提供している JST 文献検索サービス JDreamII を利用し、研究者を著者として含む論文を検索しキーワードを入手した。そして、同一研究者の論文の書誌情報のシソーラス用語および準シソーラス用語から、同一論文に付与されたすべてのキーワードのペアを抽出する。抽出されたキーワードペアのうち、共起頻度 10 以上（10 論文以上に共に付与された）キーワードペアを研究者の代表キーワードペアとする。次に、研究者の代表キーワードペアがタイトルおよび本文に含まれる特許文献を検索し、検索された特許文献の頻度を研究者と特許文献の関連度（式(2)）とする。

$$\text{Rel}(r, p) = \sum_{(kw_i, kw_j) \in K(r)}^m \text{Col}(p, kw_i, kw_j) \quad (2)$$

$$\text{Col}(p, kw_i, kw_j) = \begin{cases} 1 & kw_i, kw_j \text{ が } p \text{ で共起する} \\ 0 & \text{それ以外} \end{cases} \quad (3)$$

研究者 r と特許文献 p の関連度を $\text{Rel}(r, p)$ とする。 kw_i, kw_j は、研究者 r の学術論文書誌情報より抽出したキーワードを表し、 $K(r)$ は共起頻度 10 以上の代表キーワードペアの集合を表す。 m は共起頻度 10 以上のキーワードペアの数を表す。

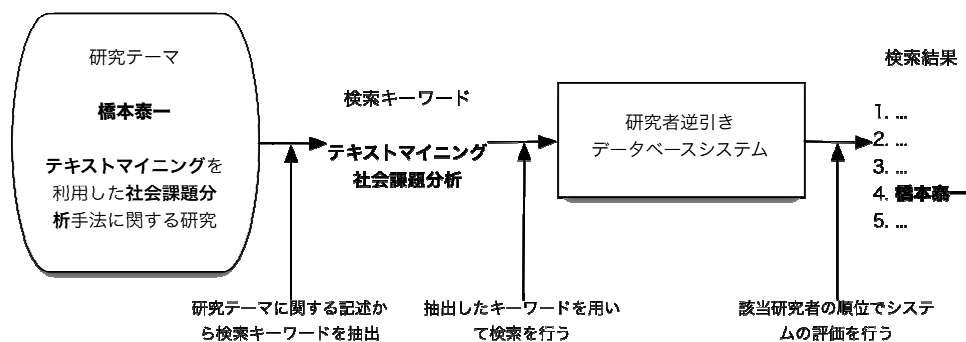


図 3 評価実験の概要

4. 評価実験

4.1. RiverStone の研究者逆引きデータベースシステムのスペック

RiverStone に構築した研究者逆引きデータベースシステムのデータについて述べる。2006 年度に東京工業大学に所属していた研究者（125 名）を対象とした。研究者-特許関連度の計算に用いた論文は約 7 万件であり、1 研究者当たりの平均論文数が約 570 件である。論文より抽出した共起頻度 10 以上のキーワードペアは約 190 万ペアであり、研究者の平均代表キーワードペアは約 1500 ペアであった。検索対象となる特許文献は 2004 年から 2007 年の公開特許公報（約 165 万件）とし、約 320 万組の研究者-特許文献の対応関係が計算され、研究者と対応付けられた特許文献はのべ約 20 万件である。

4.2. 評価実験

RiverStone に含まれる研究者（125 名）のうち、117 名の研究者が東京工業大学の HP に公開している研究テーマ（328 研究テーマ）に関する記述を用いて評価実験を行った。この研究テーマからキーワードを抽出し、そのキーワードを用いて研究者逆引きデータベースを検索する。そして、該当する研究者が検索結果の何位に現れるかによってシステムの評価を行う（図 3）。研究テーマから抽出したキーワードを用いて研究者逆引きデータベースで検索を行ったとき、システムが該当研究者を検索結果上位で出力することができれば、システムの性能がよく、そうでなければ性能が悪いと見なす。

評価実験では、検索対象となる特許文献を 2007 年のみ、2007 年から 2006 年までの 2 年間、2007 年から 2005 年までの 3 年間、2007 年から 2004 年までの 4 年間の 4 種類で検索を行い、表内の数字は該当研究者がその順位だったテーマの数を表す（表 1 および図 4）。328 の研究テーマのうち、約 120 から 140（全体の約 40%前後）の研究テーマが特許検索で検索結果を得ることができ、研究者を出力することができた。そのうち、約 70 前後（全体の約 20%前後）の研究テーマで上位 10 位以内に該当研究者を出力することができ、約 40 前後（全体の約 12%前後）で上位 2 位までに該当研究者を出力することができた。

該当研究者が出力されない、もしくは、10 位以降に検索される主な原因は、検索に用いたキーワードが様々な分野で共通で用いられる語（例：半導体、化学、合成）であり、多種多様な特許文献がヒットしてしまうためであった。また、特許文献がヒットしなかった主な原因は、検索に用いたキーワードが特許文献では使われにくい（例：材料工学、構造化学）、もしくは、検索キーワードの数が多く該当する特許文献が存在しないためであった。提案システムにおいては、特許文献の検索結果が研究者特定に重要な要素であり、特許検索結果がある共通する技術に関連したものだけにうまく絞り込むことができれば、研究者検索の結果が良くなる傾向にあった。

5. まとめ

本論文では、特許を利用した研究者逆引きデータベースの構築について報告した。我々のデータベースでは、研究者の学術論文に付与されたキーワードをもとに、研究者と特許文献の関連度を計算し、キーワードによる検索結果上位の特許と関連度が高い研究者を出力する。東京工業大学に2006年度に所属していた125名の研究者と2004年から2007年の公開特許公報との関連度を計算し、約20万件の特許に対して研究者を関連づけたデータベースを構築した。研究者の業績とは関連のない特許データベースを軸にすることにより、これまでのデータベースに比べ、より多様なキーワードによる研究者探索が行うことができるようになった。特許文献-研究者の関連度の精度向上と対応する特許文献がない場合の対応方法が今後の検討課題である。

表 1 評価実験結果 (328 研究テーマ)

	2007		2007-2006		2007-2005		2007-2004	
1 位	22/22	6.7%	25/25	7.6%	27/27	8.2%	28/28	8.5%
2 位	12/34	10.4%	11/36	11.0%	15/42	12.8%	16/44	13.4%
3 位	6/40	12.2%	4/40	12.2%	5/47	14.3%	9/53	16.1%
4 位	5/45	13.7%	7/47	14.3%	7/54	16.5%	5/58	17.7%
5 位	5/50	15.2%	7/54	16.4%	7/61	18.6%	5/63	19.2%
6 位	3/53	16.2%	4/58	17.7%	3/64	19.5%	3/66	20.1%
7 位	1/54	16.5%	4/62	18.9%	3/67	20.4%	4/70	21.3%
8 位	1/55	16.8%	2/64	19.5%	2/69	21.0%	6/76	23.1%
9 位	0/55	16.8%	5/69	21.0%	4/73	22.3%	2/78	23.8%
10 位	1/56	17.1%	1/70	21.3%	2/75	22.9%	2/80	24.4%
10 位以下	31/87	26.5%	35/105	32.0%	32/108	32.6%	35/115	35.1%
出力されない	30	9.1%	27	8.2%	25	7.6%	25	7.6%
特許文献なし	211	64.3%	196	59.8%	196	59.8%	188	57.3%

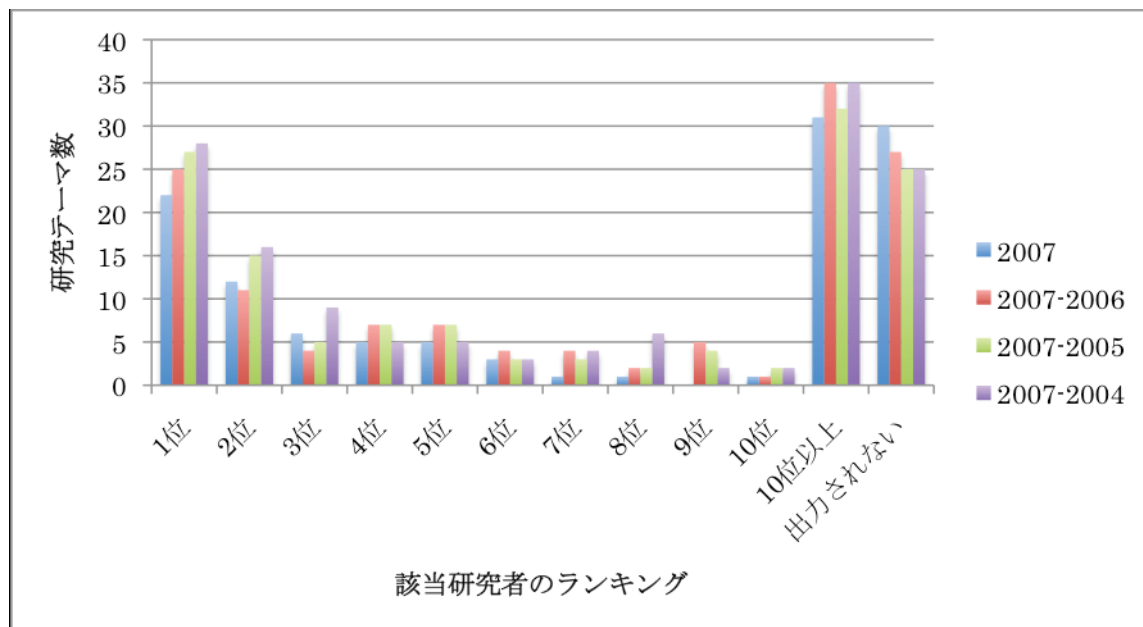


図 4 評価実験結果

謝辞

本研究は、文部科学省科学技術振興調整費「戦略的研究拠点育成プログラム」の支援の下に実施した。

参考文献

- [乾08] 乾孝司, 内海和夫, 橋本泰一, 石川正道: 新聞記事からの社会課題に対する技術的対策情報の抽出, 第7回情報科学技術フォーラム(2008)
- [橋本08a] 橋本泰一, 乾孝司, 村上浩司, 内海和夫, 石川正道: 社会課題発見のためのテキストマイニングシステム: RiverStone, 言語処理学会第14回年次大会(2008)
- [橋本08b] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道: 社会課題発見のための文書クラスタリングとクラスタ評価指標, 情報処理学会自然言語処理研究会(2008-NL-186)(2008)
- [橋本08c] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道: 文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集, Vol. 5, pp. 216-226(2008)
- [内海09] 内海和夫, 乾孝司, 橋本泰一, 村上浩司, 石川正道: 社会課題とその解決に結びつく科学技術に関する有用知識の抽出, 社会技術研究論文集, Vol. 6, pp. 187-198(2009)