研究者逆引きデータベースシステムの構築

Constructing a Reverse Researcher Database System based on the Technological Term Linkage between Researchers and Patents

橋本 泰一[*1] 乾 孝司[*2] 内海 和夫[*1] 石川 正道[*3]
Taiichi Hashimoto Takashi Inui Kazuo Utsumi Masamichi Ishikawa

[*1]東京工業大学 統合研究院
Integrated Research Institute, Tokyo Institute of Technology

[*2]筑波大学 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

[*3]東京工業大学 大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

This paper presents a new database system for discovering researchers. Our system is based on the technologial term linkage between researchers and patents. The system beforehand extracts important keywords from papers of a researcher, and calcurates relations between the researcher and patents using them. When user inputs keywords to the system, it outputs patents and researchers related them. The merit of our system is that user can discover researchers by technical keywords in patents.

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．日本国内のデータベースやサービスとしては，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 ReaD[*1]や JST 文献検索サービス JDreamII[*2]，国立情報学研究所が提供する NII 論文情報ナビゲータ CiNii[*3]や科学研究費補助金データベース KAKEN[*4]，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 SCOPUS[*7]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．前に挙げた各データベースでは，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することができる．そのため，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を代表するキーワードを抽出し，そのキーワードを多く含む特許を研究者と関連度が高い特許であると見す研究者-特許関連度を計算しておく．そして，キーワードによる特許検索を行い，検索結果の上位の特許と関連度が高い研究者をランキングし出力する．実際に，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 万件の特許に対して研究者を関連づけたデータベースを構築した．

連絡先: 橋本泰一（東京工業大学統合研究院）
〒 226-8503 神奈川県横浜市緑区長津田町 4259 S1
hashimoto@iri.titech.ac.jp

∗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/

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・開発可能な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．

3. 特許を基にした研究者逆引きデータベース

3.1 特許検索を利用した関連研究者のランキング

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と関連の深い研究を行った研究者を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

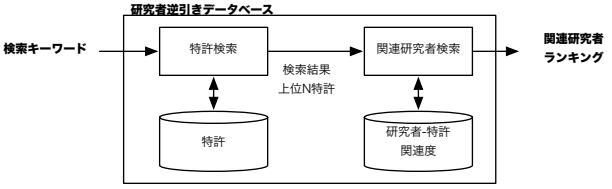

図 1: RiverStone の研究者逆引きデータベースの概要

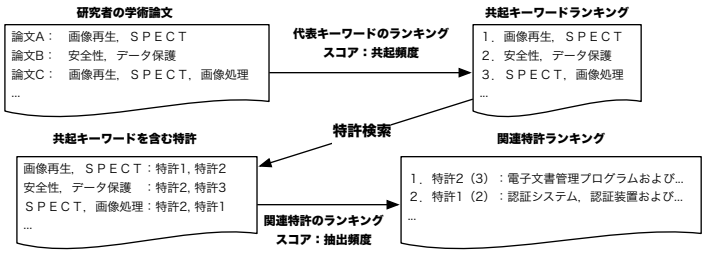

図 2: 研究者-特許関連度の計算方法

検索結果上位 n 件の特許との関連度の総和を用いる．

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

研究者 r のスコアは，式 (1) により計算される．$Rel(r, p_i)$ は，研究者 r と検索結果 i 番目の特許 p_i との関連度を表す．

3.2 研究者-特許の関連度

研究者と特許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，そのキーワードを多く含む特許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．

論文の書誌情報の入手にあたって，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 を利用した．JDreamII では，キーワードなどによる検索で該当した書誌情報をダウンロードできる．ダウンロードした書誌情報には，論文の標題，著者名，概要，シソーラス用語や準シソーラス用語[*8]などが含まれている．

まず，研究者が著者に含まれる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．そして，抽出されたキーワードペアのうち，頻度 10 以上（10 論文以上に同時に付与された）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を抽出する．そして，抽出された特許の頻度を研究者と特許の関連度（式 (2)）とする．

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

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

K は研究者 r の学術論文書誌情報より抽出した代表キーワードペアの集合を表す．

3.3 RiverStone の研究者逆引きデータベースのスペック

RiverStone に構築した研究者逆引きデータベース構築に利用した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，研究者の平均論文数が約 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 である．検索対象となる特許は 2004 年から 2007 年の公開特許公報（約 165 万件）であり，約 320 万の研究者-特許の対応関係が計算され，研究者と対応付けられた特許は約 20 万件である．

[*8] シソーラス用語とは，著者または第 3 者によって論文に付与されたキーワードのうち，科学技術振興機構が提供する「科学技術用語シソーラス」に含まれるキーワードであり，準シソーラス用語とは，「科学技術用語シソーラス」に含まれないキーワードである．

4. 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．実際に，我々が開発しているシステム RiverStone の一部として，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．

研究者-特許関連度や検索結果の評価や対象研究者の拡大が今後の課題である．

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)