様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ９ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１３７

研究課題名（和文）構造的関連性学習を用いた大規模学術情報のリンケージに関する研究

研究課題名（英文）Cross-Domain Academic Search using Structural Correspondence Learning

研究代表者
森　純一郎（Mori, Junichiro）

東京大学・工学（系）研究科（研究院）・講師

研究者番号：３０５０８９２４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：まず、数百万規模の論文情報および数十万規模の特許情報の収集を行い、これらの大規模な学術・技術テキストデータの効率的な蓄積についてデータベースの設計を行った。次に、写像関数を用いた概念空間上で、文書間の類似尺度の設計を行った。設計した類似尺度を複数の領域に適用し妥当性を検証するとともに、論文群と特許群の関連性を可視化するツールを作成し、その成果を複数の学会において発表を行った。また、「学術俯瞰システム」を構築し、実証実験を行った。実証実験においては、抽出した概念語の妥当性、線形識別器野精度、文書間の関連度計算の精度、検索結果提示の妥当性について特に評価を行った。

研究成果の概要（英文）：We propose a method to automatically associate documents from different domains such as scientific paper and patent. The proposed method enables cross-domain academic search on the scientific data. Borrowing ideas from multi-task learning and structural correspondence learning, our approach automatically identifies correspondences among the words from different domains using a small number of so-called concepts. Our method models the correlation between the concepts and all other words by training linear classifiers on the documents from different domains.

研究分野：人工知能

キーワード：計量書誌分析　構造的関連性学習　機械学習　情報検索

１．研究開始当初の背景

学術情報のオンライン化により、近年膨大な量の学術情報に容易にアクセス可能になってきている。世界最大の学術論文・文献データベースの一つであるトムソンロイター社の Web of Knowledge には現在数億万規模の学術論文・文献情報がウェブを通して検索可能になっている。国内においても、国立情報学研究所の CiNii や科学技術振興機構の J-GLOBAL といったウェブサービスを通じて、数千万規模の学術論文・文献情報が検索可能である。これらのオンラインサービスでは、学術論文・文献のみならず特許情報、企業・製品情報、研究者・研究機関情報など様々な学術情報にアクセス可能であり、学術分野の発達とウェブの普及にともない、その情報量は年々爆発的に増加している。

これらの爆発的に増加する学術情報・コンテンツに対して、異なる学術情報を結びつける「リンケージ」という処理が注目されている。特に、論文と特許の間のリンケージは「サイエンスリンージ」と呼ばれ、学術研究がどのように製品やサービスに寄与しているかを示す指標となっている。サイエンスリンケージは、基本的には特許が引用する論文の数によって算出される事が多いが、近年膨大な学術情報のテキストが分析可能になってきたことにより、膨大なテキスト情報を分析し、論文、特許、製品といった複数の関連した学術情報を横断的にリンケージすることが着目されている。このようなリンケージは膨大な学術情報の検索向上へもつながる。しかしながら、語彙や文脈が異なるため、表層的な関連付けが困難な異種学術情報をどのように結びつけるかは、膨大な学術情報のリンケージを実現する上で大きな課題である。本研究では、機械学習手法である構造的関連性学習を用いた大規模学術情報のリンケージに関する研究を行う。

２．研究の目的

本研究では、構造的関連性学習を用いた大規模学術情報のリンケージ手法について具体的に以下の項目について研究を行う。

(1)　構造的関連性学習を用いて異種の学術情報間の関連性を計算する技術

(2)　任意の入力文書に対して異種の学術情報を横断して文書検索を行う技術

まず(1)については、従来の文書類似尺度では関連付けが難しい大規模な異種の学術情報の文書間をどのように関連づければよいのかを明らかにする。そのためにまず、国立情報学研究所や科学技術振興機構など学術情報を保持する機関と連携し、論文、特許、企業、ウェブ情報の異なる情報を対象に大規模な学術データの収集を行った上で、これら異なる情報の文書を関連づける手法について明らかにする。具体的には、教師なしの学習である構造的関連性学習により、高次の概念空間に文書を写像することで文書間の潜在的な関連性計算手法の設計・実装を行う。

次に(2)については、(1)で研究開発を行った技術を元に、実際の情報検索システムにおいて、どのように異種の学術情報を横断した文書検索を行えばよいのかを明らかにする。そのために、任意の入力文書（例えば論文）に対して、異なる学術情報（特許、企業、ウェブ）の関連する文書を検索して提示するシステムを構築し、従来の学術情報検索サービスと連携した実証実験を通じて、異種学術情報を横断した文書検索手法について明らかにする。具体的には、入力文書に対して異なる学術情報の文書をランキングする手法および効率的に検索結果を提示する手法の設計・実装を行う。

本研究で行う異なる学術情報のリンケージについて関連するのは、論文と特許とつながりを見るサイエンスリンケージであるが、サイエンスリンケージは基本的に引用情報に基づく元でありテキスト情報を扱っていない。本研究で用いる構造的関連性学習は、近年異なる領域・分野のテキストの潜在的な関連性を抽出する手法と着目されており、多言語文書間の関連抽出、異なる商品のレビュー間の関連抽出などにおいて有効性が示されてる。しかしながら、本研究の対象とする膨大な異種学術情報の関連性抽出に適用している既存研究はなく、構造的関連性学習の異種学術情報への応用とそこから得られる知見ならびシステムの構築は本研究の特色である。

本研究の成果として得られる大規模学術情報のリンケージ手法により、爆発的に増加する論文、特許、企業・製品、ウェブといった膨大な学術情報を横断的に関連づけることが可能になり、従来の学術検索サービスの向上への貢献が期待できる。特に、本研究で構築する異種学術情報検索システムから得られる知見により、大規模学術情報検索サービスの設計の指針を得ることが期待できる。また、本研究の成果は学術情報のみならず異種のテキスト情報を関連づける汎用的な手法であり、膨大なウェブ文書、企業内文書、政府官公庁の文書などを横断的に関連づけることで一般的な情報検索にも寄与することが期待できる。

３．研究の方法

本研究では、構造的関連性学習を用いた大規模学術情報のリンケージについて、次の 2 つの主たる技術(1),(2)を大目標として研究開発を行う。各技術は具体的に以下に示す研究項目を行うことで実現を目指す。

(1) 構造的関連性学習を用いて異種の学術情報間の関連性を計算する技術

①　大規模な異種学術情報の収集と特徴抽出手法の設計と実装

②　構造的関連性学習により学術情報の文書を高次概念空間へ写像する手法の設計と実装

③　高次概念空間において文書間の関連性を計算する手法の設計と実装

(2) 任意の入力文書に対して異種の学術情報を横断して文書検索を行う技術

④　論文・特許・企業・ウェブの学術情報を対象とした検索システムの設計と実装

⑤　異種学術情報検索結果のランキングと提示手法の設計と実装

⑥　実証実験による手法およびシステムの評価および改善

各研究項目について研究実施の各年度に次のように研究を進める。

①大規模な異種学術情報の収集と特徴抽出手法の設計と実装

大規模学術情報として主に論文（研究者）情報、特許情報、企業（製品）情報を対象にデータの収集を行う。論文データの収集にあたっては、NII 国立情報学研究所の論文データベース CiNii および研究者データベース KAKEN が提供する API を利用して数百万規模の論文および研究者情報を効率的に収集する。また特許データの収集にあたっては、JST 科学技術振興機構の科学技術情報ポータル J-GLOBAL が提供する API を利用して数十万規模の特許情報を効率的に収集する。また J-GLOBAL からは科学技術用語データも収集し文書の特徴抽出に利用する。企業データの収集にあたっては、中小企業庁、中小企業基盤整備機構ならび各地域の産学連携支援機関と連携して私が研究開発を行っている企業情報検索システムSMEETのデータを活用して数万規模の企業情報を効率的に収集する。収集したデータは検索可能なように文書処理を行い高速なデータベースに蓄積する。各データの文書処理にあたっては、文書を特徴づけるキーワードの抽出を行う。キーワードは科学技術用語データの語彙情報およびデータの中の頻度・共起の統計情報に基づいてスコア付けを行う。各文書をキーワードのベクトルとして表現することで以後の処理を行う。

②構造的関連性学習により学術情報の文書を高次概念空間へ写像する手法の設計と実装

収集したデータを分析し、任意の異なる学術情報間（論文データと特許データ、論文データと企業データなど）に共通して出現する少数のキーワードを「概念語」群として抽出する。概念語の抽出にあたっては事前に収集した用語・類義語データを活用する。次に、概念語とその他のキーワードの関連性のモデルを構築するため、テキスト中のキーワード群から概念語の出現を予測するような線形分類器を学習する。異なる学術情報を横断する線形分類器を学習するため、任意の異なる学術情報データの文書を組み合わせることで大規模な学習データを構築する。次に、学習により得られた線形分類器の重み行列の次元を削減することで、任意の文書のキーワードベクトルを高次の抽象的な「概念」の空間に写像する関数を導出する。

③高次概念空間において文書間の関連性を計算する手法の設計と実装

異なる学術情報のテキストデータを高次の概念空間に写像することで、元のデータ空間では関連性の薄い文書同士を、潜在的な概念を通して関連付けを行う。文書間の関連度は、各文書を概念のベクトルとして、複数のベクトル間類似尺度を適用しタスクに応じて最適な類似尺度の選択を行う。また、複数の類似尺度の出力を集約する尺度の設計・実装を行う。

④論文・特許・企業・ウェブの学術情報を対象とした検索システムの設計と実装

平成 24 年に設計・実装を行った異種の学術情報の文書間の関連性を計算する技術を元に、学術情報検索システムの設計と実装を行う。同システムの入力となるのは、任意の学術情報の文書であり、例えば、ある論文をシステムを入力するとその論文に関連する特許情報、企業情報およびウェブ情報を出力として提示する。また、同機能を提供する API の設計と実装も行い。システムが容易に他のシステムと連携可能なようにする。

⑤異種学術情報検索結果のランキングと提示手法の設計と実装

検索システムにおいて、入力文書に関連する異なる学術情報の文書の一覧を提示際、その検索結果において複数の学術情報を適切に混在させランキングするための手法の設計と実装を行う。ランキングは基本的には文書間関連度のスコアに基づくが、後述する実証実験によりユーザのフィードバックを得ることで関連度計算手法および検索結果の提示手法の改善を行う。特に検索結果提示手法についてはユーザインタフェースの専門家と協力する。

⑥実証実験による手法およびシステムの評価および改善ならびに成果公開

上記により設計・実装を行った異種学術情報検索システムについて平成 25 年度下半期においてシステムを公開し実証実験を実施する。実験においては、私が研究開発してきた大規模論文分析システム、企業情報検索システムの他、国立情報学研究所の CiNii や科学技術振興機構の J-GLOBAL といった既存の学術情報検索システムとも連携することで、さまざまなユーザを対象に異種学術情報の横断検索機能を提供し、大規模な利用統計情報を蓄積する。実証実験においては、抽出した概念語の妥当性、線形識別器の精度、文書間の関連度計算の精度、検索結果提示の妥当性について特に評価を行い、その結果をもとに再度手法およびシステムの改善を行う。

４．研究成果

(1)平成 24 年度

まず、論文データベースの Web of Science を対象に数百万規模の論文情報の収集を行った。また、特許データベースの Thomson Innovation を対象に、数十万規模の特許情報の収集を行った。あわせて、収集を行ったこれらの大規模な学術・技術テキストデータの効率的な蓄積についてデータベースの設計を行った。

次に、収集した大規模な学術・技術テキストデータを分析し、分野・階層横断的に使用される「概念語」の抽出を行うための重み付け手法の研究開発を行った。また、概念語とその他の語の関連性のモデルを構築するため、テキスト中の語群から概念語の出現を予測するような大規模な線形分類器の学習を行った。さらに、学習により得られた線形分類器の重み行列の次元を削減することで、任意の文書ベクトルを高次の抽象的な「概念」の空間に写像する関数の設計を行った。

最後に、写像関数を用いた概念空間上で、文書間の類似尺度の設計を行った。設計した類似尺度を複数の領域に適用し妥当性を検証するとともに、論文群と特許群の関連性を可視化するツールを作成し、その成果を複数の学会において発表を行った。

(2)平成 25 年度

平成 24 年度に設計・実装を行った異種の学術情報の文書間の関連性を計算する技術を元に、学術情報検索システムの設計と実装を行った。同システムの入力となるのは、任意の学術情報の文書であり、例えば、ある論文をシステムに入力するとその論文に関連する特許情報、企業情報およびウェブ情報を出力として提示するものである。

次に、検索システムにおいて、入力文書に関連する異なる学術情報の文書一覧を提示する際、その検索結果において複数の学術情報を適切に混在させランキングするための手法の設計と実装を行った。

最後に、上記により設計・実装を行った異種学術情報検索システムについて、「学術俯瞰システム」を構築し、実証実験を行った。実証実験においては、抽出した概念語の妥当性、線形識別器野精度、文書間の関連度計算の精度、検索結果提示の妥当性について特に評価を行い、その結果をもとに再度手法およびシステムの改善を行った。

(3)平成 26 年度

平成 25 年度の学術情報検索システム研究開発および実装において、構造性関連性学習を用いた大規模学術情報のリンケージ手法について、ニューラルネットワーク言語モデルと組み合わせることにより、精度向上が図れるという知見を得たために、引き続き手法の改善と評価を行った。これらの研究成果はウェブサービスとして「学術俯瞰システム」として、実証実験後も広く一般に利用可能な形で公開している。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Shino Iwami, Junichiro Mori, Yuya Kajikawa and Ichiro Sakata, “Detection method of emerging leading papers using time transition”, Scientmetrics, Vol. 101, 2014, pp. 1515-1533

② Katsuhide Fujita,Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata,Detecting Research Fronts Using Different Types of Weighted Citation Networks, Journal of Engineering and Technology Management, Vol. 32, 2014, pp. 129-146

〔学会発表〕（計 11 件）

① Juniki Marui, Nozimi Nori, Takeshi Sakaki, and Junichiro Mori, “Empirical Study of Conversational Community using Linguistic Expression and Profile Information”, The 2014 International Conference on Active Media Technology (AMT2014), 2014 年 8 月 11 日-8 月 14 日, Warsaw, Poland

② Shino Iwami, Junichiro Mori, Yuya Kajikawa and Ichiro Sakata, “Bibliometric Methodology to Detect Collaborative and Competitive Countries”, The IEEE International Conference on Industrial Engineering and Engineering Management, 2014 年 12 月 9 日-12 月 12 日, Kuala Lumpur, Malaysia

③ Shino Iwami, Junichiro Mori, Yuya Kajikawa and Ichiro Sakata, “Comparison of Indicators to Detect Emerging Researches using Time Transition in Quasicrystals”, The IEEE International Conference on Industrial Engineering and Engineering Management, 2013 年 12 月 10 日-12 月 13 日, Bangkok, Thailand

④ Shino Iwami, Junichiro Mori, Yuya Kajikawa and Ichiro Sakata, “Detection of Next Researches using Time Transition in Fluorescent Proteins”, 14th International

Conference on Scientometrics and Informetrics (ISSI2013), 2013年7月15日-7月19日, Vienna, Austria

⑤ 関喜史、森純一郎、ウェブを用いたサイエンスマップのためのクエリ拡張の研究、人工知能学会全国大会、2013年6月4日-6月7日, 富山県

⑥ Shino Iwami, Junichiro Mori, Yuya Kajikawa, T. Uehara, and Ichiro Sakata, Detection of Promising Fields using Time Transitions in Cryptology, 22nd International Conference for Management of Technology" (IAMOT2013), 2013年4月14日-4月18日, Porto Alegre, Brazil

⑦ Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata, Detecting Research Fronts using Citation Network Analysis, 2012 Annual Meeting of Institute for Operations Research and Management Sciences (INFORMS2012), 2012年10月14日-10月17日, Arizona, USA

⑧ Katushide Fujita, Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata, Detecting Research Fronts Using Different Types of Combinational Citation, 17th International Conference on Science and Technology Indicators (STI 2012), 2012年9月5日-9月8日, Quebec, Canada

⑨ Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Detecting research fronts using different types of weighted citation networks, Proc. Portland International Center for Management of Engineering and Technology 2012 (PICMET), 2012年7月29日-8月2日, Vancouver, Canada

⑩ Vitavin Ittipanuvat, Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Finding linkage between technology and social issue: A literature based discovery approach, Proc. Portland International Center for Management of Engineering and Technology 2012 (PICMET), 2012年7月29日-8月2日, Vancouver, Canada

⑪ Vitavin Ittipanuvat, Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Measuring relatedness between technology and social issue citation networks, Proc. International Society for Professional Innovation Management (ISPIM), 2012年6月17日-6月20日, Barcelona, Spain

〔図書〕（計0件）

〔産業財産権〕
○出願状況（計0件）

○取得状況（計0件）

〔その他〕
ホームページ等
学術俯瞰システム
http://academic-landscape.com

６．研究組織
(1)研究代表者
森　純一郎（MORI, Junichiro）
東京大学・大学院工学系研究科・特任講師
研究者番号：３０５０８９２４