様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １８ 日現在

機関番号：３７１０９
研究種目：基盤研究(C)（一般）
研究期間：2018～2023
課題番号：１８Ｋ１１９９０

研究課題名（和文）論文の細粒度情報を用いた研究の相互関係理解

研究課題名（英文）Understanding the Interrelationships of Studies Using Fine-Grained Information in Papers

研究代表者
中藤　哲也（Nakatoh, Tetsuya）

中村学園大学・栄養科学部・准教授

研究者番号：２０２５３５０２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：学術論文の評価指標に関する研究を進めた。書誌情報と論文の引用数に関係があることを機械学習により明らかにした。また、新しい指標「引用グループ数」を提案し、その有効性を示した。さらに Focused Citation Count(FCC)およびRevised Focused Citation Count(RFCC)を提案し、引用数(CC)よりも高い精度で評価できることを確認した。論文をセクション単位に自動分割するシステムを構築し、セクション間の類似度を定義し、引用元セクションを自動判別することで、論文間の関係を詳細に可視化するシステムの開発を進めた。

研究成果の学術的意義や社会的意義
関連研究の調査は研究者にとって重要です。新しい研究を始める際には、その新規性を確認する必要があります。また、研究成果を公開する際には、関連する他の研究に対して自分の研究を位置づけ、差別化することが求められます。しかし、膨大な論文の中から適切なものを見つけ、その内容を理解して関連性や違いをまとめることは非常に時間のかかる作業です。本研究の成果により、学術論文の分析が効率化され、学術研究の推進・発展に寄与することが期待されます。

研究成果の概要（英文）：We conducted research on evaluation metrics for academic papers. Using machine learning, we demonstrated a relationship between bibliographic information and the number of citations a paper receives. Additionally, we proposed a new metric called "Citation Group Count" and showed its effectiveness. Furthermore, we introduced the Focused Citation Count (FCC) and the Revised Focused Citation Count (RFCC), confirming that they provide higher accuracy in evaluation compared to the traditional Citation Count (CC). We also developed a system to automatically divide papers into sections, define similarities between sections, and automatically identify the citing sections, thereby advancing the development of a system to visualize the relationships between papers in detail.

研究分野：情報科学

キーワード：情報抽出　計量書誌学　データベース　テキストマイニング

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

関連研究の調査は研究者にとって非常に重要なタスクである。新たな着想に基づいて研究をスタートさせる際には、その研究の新規性を見定める必要がある。また、新しい研究成果が得られ、それを論文等で公開する際にも、関連する他の研究に対して自らの研究を位置付け、あるいは差別化する事が求められる。しかしながら、膨大な研究論文の中から、適切かつ重要な論文を見つけ出し、その内容を理解し、研究内容の関連性や違いを文章にまとめる事は容易ではなく、非常に時間のかかる作業となる。更に、学術領域の細分化や学際領域研究の増加などから、関連する研究を見落とすことなく見つけることが、より困難になっている現状がある。
多くの学術論文データベースは使い易さの改善と機能向上を図っており、研究に関するキーワードを用いた検索をおこなう事によって、関連する研究のリストアップは容易になった。しかし、単純なキーワードマッチングによる検索は、目を通す必要があるかも知れない大量の論文をリストアップし、それだけの数の論文をどの程度の深さで読み込むのか、という新しい問題を研究者に投げかけている。引用数などの評価指標によるランキングもあるが、チェックすべき論文を絞り込むには充分とは言えない。結局、研究者は膨大な論文を一遍ずつ読み込んで理解した上で、自らの研究との関係を見出さなければならない。
このように学術研究における関連研究の調査には、情報処理技術・人工知能技術を用いた研究調査アシストが必要であると考える。関連研究の調査が完了していない論文草稿、あるいは調査を行いたい研究に関する文書を、テキストマイニング、機械学習などの技術を用いて分析し、様々な観点（例えば、目的、手法、材料など）において密接に関連する研究論文をデータベースから抽出し、それらの論文同士の関係を自動的に分析する。研究者はより客観的な関連研究の調査を半自動的に行うことができ、より詳細な比較検討が必要だと感じた関連研究のみに集中することが可能となる。客観的なデータに基く関連研究調査は論文の質の向上を可能とし、半自動的な調査は論文の生産性を上げる事が可能となる。

２．研究の目的

関連研究の調査を自動的に行う事を目的とし、次の 2 点を学術的「問い」として挙げる。
・研究者は、どのような場合に研究が関連していると判断するのか？
・研究者は、何にどのように着目して、２つの研究の違いを理解するのか？
これらは研究者が普通に行っている作業であるが、本研究課題ではこれらの「問い」に対する「答え」の明文化に取り組む。この「答え」は研究者自身が明確に認識しているものではない。このため、本研究では分析・調査と検証を繰り返す事により「答え」に近付くことを試みる。
この「問い」に答える事を最終目標にし、以下の点を本研究における具体的な目的とする。
(A) 与えた研究内容に関する関連研究を見つけ、それらの関係を分析することにより、
(B) ２論文の関係を明確に示す文章を生成する。また、
(C) 複数の論文の関係をひと目で理解できるグラフを生成する。

３．研究の方法

関連研究の調査が完了していない論文草稿、あるいは調査を行いたい研究に関する文書を、テキストマイニング、機械学習などの技術を用いて分析する。それにより、その研究内容に関する関連研究を見つけ、２論文の関係を明確にする。これを達成するために、次の手順で研究を実施する。
１：(a)学術論文をデータベース等から収集し、分析用データベースに登録する。(b)複数の手法を用いて論文を特徴ベクトル化する。(c)分析対象論文を細粒度「情報の単位を論文全体、Section 毎、文 n-gram、単文、単語など」に分割し、(b)で求めた異なるタイプの特徴ベクトルを用い、機械学習により引用-被引用の関係の有無を分類する。どの情報粒度、特徴ベクトル化タイプが適切であるかを評価する。(d) 研究者を対象とした聞き取り調査を行う。研究者本人が、何にどのように着目して他の論文の引用を行ったのかをデータベース化する。聞き取り調査の設計後、アルバイト等を利用して多くの情報を収集する。２：(e) (c)の結果と(d)のデータを比較し、研究者のどんな視点が実際のデータに現れるのかを分析する。(c)の精度が低い場合、(d)の調査結果を踏まえた論文データの特徴ベクトル化を検討する。(f) 引用関係の分類に効いているデータから単語や表現を抜き出し、それらが２つの論文の何を説明しているのかを分析する。分析結果を用いて、論文の関係を表す説明文の自動生成を試みる。上手く行かない場合はどのような情報が不足しているかを検討し、(b)及び(c)からそれらの抽出を試みる。平成 32 年度：(g) (c)と(f)から論文同士の関係の可視化を行う。特に情報の粒度を変更する事によって論文同士の関係が変化する事を示し、様々な観点からの関連研究の分析が行えるように構築する。

４．研究成果

まず、学術論文の評価指標に関する研究を中心に推進した。学術論文の客観的評価に関する研究として、アブストラクトを含む学術論文の書誌情報と該当論文の引用数の関係を分析した。研究対象のデータとして(1)コンピュータサイエンス分野の著名な論文誌及びプロシーディングの計 20 誌から抽出した 38,766 件の論文の書誌情報、(2)Bibliometrics(計量書誌学)分野の論文として論文データベースから抽出した 10,186 件の論文の書誌情報を準備した。これらの論文の書誌情報を素性とした特徴ベクトルを生成し、機械学習によって引用数で論文を分類する分類器を生成した。分類する引用数の閾値を変化させた大量の分類器の性能評価から、書誌情報と引用数に一定の関係が存在することを示した。この情報を用いることで学術論文の適切な選択や、学術論文間の関係抽出の精度を改善できる。
次に、引用数ではなく引用グループ数を用いた論文評価手法についての研究を行った。上記(2)の 10,186 件の論文を引用している論文に関して著者グループを構築し、どの著者グループがどの論文を引用しているかについて分析を行い、特定の研究グループによる多数の引用に比べて、多くの研究グループによる引用は該当論文の影響の広さを示す指標となりうることを示した。引用数などの他の評価指標と組み合わせることで、学術論文の多面的な評価を行うことが可能となる。
一般的に用いられている引用数 Citation Count(CC)に対して、より特定の注目分野に特化した基準による学術論文の評価方法として Focused Citation Count (FCC)及び Revised Focused Citation Count (RFCC)を提案した。Bibliometrics(計量書誌学)分野の論文データベースから抽出した計 10,186 件の論文の書誌情報のデータベースに対して、これらの引用数の評価手法を適応し、被験者による事前判定を解とした精度 Presicion@N を求め、評価を行った。結果、RFCC は CC よりも高い精度が得られることが明らかになった。また、質的評価においては、RFCC は FCC よりも的確に特定の関連分野の論文を見つけ出すことも明らかになった。
継続して論文データの拡充を行った。収集済みの Bibliometrics（計量書誌学）分野の計 10,186 件の論文の書誌情報データに加え、Scopus データベースから研究費助成金【日本学術振興会の科学研究費助成金、及びアメリカ合衆国の National Science Foundation(NSF)】を受けているコンピューターサイエンス（情報工学）分野の論文 11 年分の書誌情報データ、計 75,482 件の書誌情報データを収集し、更にこれらの書誌情報データを用いて収集可能なものに関して該当論文の全文の取得を行った。また、Pubmed より抽出した 1975 年から 2024 年までの遠隔医療に関する論文 116,740 件も収集した。これまでに全文を取得できた論文に関しては、セクション単位（基本とされる Introduction, Materials and Methods, Results and Discussion の各セクションに加え、Abstract セクション、Conclusion セクション）のパーツに自動的に分割するシステムを構築した。さらに各セクションのテキスト情報に関する分析によりその統計的性質を明らかにした。論文のセクション間の類似度を複数の手法で定義し、論文に記載された関連研究のリファレンス情報から得られる他の論文への引用に関して、リファレンスの情報には含まれていない「該当論文中のどのセクションに対する引用」なのかを自動的に判別し、より詳細に論文間の関係を明示することができる可視化システムの構築を試みているが、分析可能な論文データが限られており、明確な可視化には至っていない。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Ishita Emi、Nakatoh Tetsuya	４．巻 13636
２．論文標題 Differences Between Research Projects in Computer Science Funded by Japanese and American Agencies	５．発行年 2022年
３．雑誌名 Lecture Notes in Computer Science (Proc. of International Conference on Asian Digital Libraries 2022)	６．最初と最後の頁 144―152
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-21756-2_12	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

１．著者名 Nakatoh Tetsuya、Kodama Hironori、Hori Yuko、Ishita Emi	４．巻 13133
２．論文標題 Enriching the Metadata of a Digital Collection to Enhance Accessibility: A Case Study at Practice in Kyushu University Library, Japan	５．発行年 2021年
３．雑誌名 LNCS	６．最初と最後の頁 411～418
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-91669-5_32	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

１．著者名 Tetsuya Nakatoh, Takahiko Suzuki, Tsukasa Kamimasu, Sachio Hirokawa	４．巻 6(2)
２．論文標題 Detection of Unnatural Parts of Statistical Data	５．発行年 2020年
３．雑誌名 Information Engineering Express	６．最初と最後の頁 20-36
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

１．著者名 Zeng Chao、Nakatoh Tetsuya、Hirokawa Sachio、Eguchi Masanari	４．巻 14
２．論文標題 Text mining of tourism preference in a multilingual site	５．発行年 2019年
３．雑誌名 IEEJ Transactions on Electrical and Electronic Engineering	６．最初と最後の頁 590～596
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/tee.22841	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

1．著者名 Nakatoh Tetsuya、Hirokawa Sachio	4．巻 11569
2．論文標題 Evaluation Index to Find Relevant Papers: Improvement of Focused Citation Count	5．発行年 2019年
3．雑誌名 Lecture Notes in Computer Science (LNCS)	6．最初と最後の頁 555～566
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-22660-2_41	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Nakatoh Tetsuya、Hirokawa Sachio、Minami Toshiro、Nanri Takeshi、Funamori Miho	4．巻 23(2)
2．論文標題 Attribute-based quality classification of academic papers	5．発行年 2018年
3．雑誌名 Artificial Life and Robotics	6．最初と最後の頁 235～240
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10015-017-0412-z	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　2件／うち国際学会　5件）

1．発表者名 Tetsuya Nakatoh and Sachio Hirokawa
2．発表標題 Counting Research Groups for Citation Assessment
3．学会等名 ICADL2018（国際学会）
4．発表年 2018年

1．発表者名 Tetsuya Nakatoh
2．発表標題 Quality classification of academic papers using the attributes
3．学会等名 SCML2019（招待講演）（国際学会）
4．発表年 2019年

１．発表者名
Yasuhiro Yamada and Tetsuya Nakatoh

２．発表標題
Tag Recommendation for Open Government Data by Multi-label Classification and Particular Noun Phrase Extraction

３．学会等名
10th International Conference on Knowledge Management and Information Sharing（国際学会）

４．発表年
2018年

１．発表者名
Takahiko Suzuki, Tsukasa Kamimasu, Tetsuya Nakatoh, and Sachio Hirokawa

２．発表標題
Identification of Unnatural Subsets in Statistical Data

３．学会等名
IIAI AAI 2018（国際学会）

４．発表年
2018年

１．発表者名
Tetsuya Nakatoh

２．発表標題
Current State of Open Data in References

３．学会等名
AITIA 2024（招待講演）（国際学会）

４．発表年
2024年

１．発表者名
中藤 哲也，石田 栄美

２．発表標題
オープンデータとオープンアクセスの現状分析について

３．学会等名
情報処理学会 第86回全国大会

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	廣川　佐千男 (Hirokawa Sachio) (40126785)	東京都立産業技術大学院大学・産業技術研究科・研究員 (22605)	
研究分担者	石田　栄美 (Ishita Emi) (50364815)	九州大学・データ駆動イノベーション推進本部・教授 (17102)	
研究分担者	鈴木　孝彦 (Suzuki Takahiko) (90243906)	九州大学・情報基盤研究開発センター・准教授 (17102)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関