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　２ 年　 ６ 月　 ７ 日現在

機関番号：１７１０２
研究種目：基盤研究(A)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０１７２１

研究課題名（和文）ユーザーの視点に立った高度な学術論文検索支援に関する総合的研究

研究課題名（英文）Comprehensive Research on Advanced Support for Searching Academic Papers from the User's Perspective

研究代表者
冨浦　洋一（Tomiura, Yoichi）

九州大学・システム情報科学研究院・教授

研究者番号：１０２１７５２３

交付決定額（研究期間全体）：（直接経費）　　30,200,000円

研究成果の概要（和文）：研究の新規性の確認のための論文検索では，情報要求に関連する論文を極力漏れなく検索できることが望ましい．トピック分析結果を利用して，従来手法より検索漏れが少ない検索手法を開発した．また，検索時に有用なキーワードや日英の対訳専門用語対等の情報の抽出手法と類似文献の推薦手法を開発し，さらに，特定分野における技術動向を俯瞰する手法や実験に用いたデータ等が公開されているURLを論文中から抽出して学術資源リポジトリを自動構築する手法を提案した．
研究歴や検索の目的に応じて，どのようなデータベースを用いてどのような情報を基に読むべき論文を絞り込んでいるのかを調査し，図書館における人的支援を検討した．

研究成果の学術的意義や社会的意義
これまで，検索システムの評価は，ランキング上位の文献の内の情報要求に関連する文献の割合が重視されてきた．ランキングのある程度下位まで見た場合の検索漏れが少ないことを指標として検索手法を開発・評価した研究はほとんどない．また，論文から抽出されたキーワードや対訳専門用語対は検索時に有用であり，自動構築される学術資源リポジトリはオープンデータにも貢献する．さらに，研究歴や検索の目的に応じて，どのようなデータベースを用いてどのような情報を基に読むべき論文を絞り込んでいるのかを体系的に調査した研究はなく，調査結果は図書館における人的支援検討に役立つだけでなく，論文検索システムの新たな機能開発に繋がる．

研究成果の概要（英文）：When searching academic articles to confirm the novelty of your research, it is desirable to be able to search articles related to your research without omission. We developed a search method using topic analysis, by which the related articles missed are less than by the conventional methods. In addition, we developed methods for extracting information such as keywords and Japanese-English bilingual technical term pairs that are useful for searching and a method for recommending similar articles. Furthermore, we proposed a method for a quick overview of technological trends in specific fields, and a method for automatically constructing an academic resource repository by extracting the URL where the data used in the experiment is published from the articles.
We also investigated what kind of databases are used to narrow down the articles to be read based on what kind of information by various researchers with various purpose of search, and examined the human support in the library.

研究分野：自然言語処理，図書館情報学

キーワード：学術論文　検索支援　トピック分析　網羅性　検索行動分析　対訳専門用語　学術リソースリポジトリ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究の推進および学術論文の生産性が求められる研究者にとって，効率よく学術情報の検索ができるシステムは極めて重要である．「知」の資産を創出し続け，科学技術を文化として育む国を目指す我が国にとっても，このような検索技術の開発の意義は大きく，第４期科学技術基本計画等でも，研究情報基盤の重要さが指摘され，学術情報検索機能の強化の必要性が主張されている．

Google 等の Web 検索エンジンを用いた日常的な検索では，多くの場合網羅的に情報を収集する必要はないため，検索語による検索で十分情報要求を満たす検索が可能である．一方，学術論文検索では，多くの場合，網羅的に情報を収集する必要がある．特に，研究の新規性を確認するための学術論文検索の場合，情報要求に関連する論文を網羅的に収集できることが重要である．

学術論文検索の問題は，検索システムの機能の問題だけでなく，それを使う利用者の方にも問題がある．学生や若手の研究者は，学術論文検索の際に，学術データベースの検索機能を十分に利用していない可能性がある．また，逆に，データベースから電子ジャーナルへのリンクが一般的となった現在，以前には行われていなかった効率的な検索をしている研究者がいる可能性もある．

２．研究の目的

学術論文の検索では，検索結果の網羅性が要求されるが，情報要求を検索者自身が言語化した詳細な検索語では，関連する重要な論文を漏らしてしまう恐れがある．したがって，本研究では，情報要求に関連する可能性がある論文を網羅的に収集する検索手法を開発する．また，通常の論文調査では，表題等を利用して検索結果の論文を数百の論文に絞り込んだ後，抄録を確認し，新規性の確認や研究動向調査といった論文調査の目的に応じて論文を選択するという負荷の高い作業を行う．この負荷を軽減するための支援に利用する情報として，抄録等から読むべき論文を特定するのに有効な情報や研究動向を俯瞰的に見るために有効な情報を抽出する手法を開発する．

また，図書館における若手研究者への人的支援や検索システムの新たな機能を検討するために，修士課程学生を含む研究者が学術論文検索の際に読むべき論文をどのようにして選択しているのかの傾向を分析する．

３．研究の方法

(1) 情報要求に関連する可能性がある論文を網羅的に収集する検索手法の開発

Latent Dirichlet Allocation (LDA)を用いたトピック分析を利用する．LDA によるトピック分析では，多くの文書で共起する語のグループが１つのトピックを構成する（１つのトピックで比較的高い確率で出現する）と分析される．語 w と w’がほぼ同じことを表す別表現だとすると，w とよく共起する語のグループと w’とよく共起する語のグループは類似していると考えられ，LDA による分析において，w と w’は同じトピックが付与されると期待される．そこで，検索対象の論文集合（抄録集合）を LDA を用いてトピック分析し，その結果に基づいてクエリ中の検索語もトピックの AND/OR に変換し，トピックレベルでこれを満たす抄録がクエリを満たすと考え，このようなトピックベースのブーレアン検索を実装した．ただし，LDA によるトピック分析では，トピック数などの LDA のモデルのパラメタの値を指定して分析が行われるが，適切なパラメタ値は，トピック分析の対象の文書集合によって異なる．しかし，自動的に最適なパラメタ値を求めるのは困難であり．検索者にその設定を行わせるのも無理である．一方で，トピックベースのブーレアン検索に対する予備実験を通して，パラメタを細かく調整したとしても，高い確率で同一のトピックから生成される単語グループおよび高い割合で共通のトピックを含む文書グループの大まかな傾向は変わらない可能性があることが分かった．そこで，本研究では，検索対象の文書集合ごとに最適なパラメタ値を設定する代わりに，１つの抄録集合に対して複数のトピック分析結果を取得し，それぞれの分析結果を用いてトピックベースのブーレアン検索を実行し，この複数の検索結果に含まれる数で個々の文書をランキングする手法を開発した（この検索を本研究では Topic Search と呼ぶ）．ランキング上位の抄録は，様々なパラメタ設定に対する検索結果の多くに出現する論文であり，安定的な単語および抄録間の関係に基づいた関連論文と考えられる．

また，予備的な評価実験の際に，Topic Search と Wei らの検索モデル（Wei, X. and Croft, W.B.: LDA-based Document Models for Ad-hoc Retrieval, Proc. SIGIR, pp.178-185 (2006)）のランキングの傾向が異なることに気付いた．ある検索課題に対する Topic Search と Wei らの検索モデルそれぞれによる検索結果に対して，ランキングの 5% 範囲（上位 5%，上位 5～10%，上位 10～15%，…）の検索結果の precision と両手法による検索結果の重複率を図１に示した．どちらの手法による結果も，ランキング上位 10%までには，関連論文が比較的多く含まれていることから，それぞれのモデルにおいて高くランク付けされている論文は関連論文である可能性が高いといえる．一方で，論文の重複率は，上位 5%および上位 5～10%においてそれぞれ 0.287，0.147 と比較的低いことから，それぞれの検索モデルにおけるランキング結果の傾向は大きく異なる可能性が高い．そのため，２種類の検索モデルによるランキング結果を統合することで，それぞれのランキング結果とは異なるより良いリランキング結果が得られると期待できる．

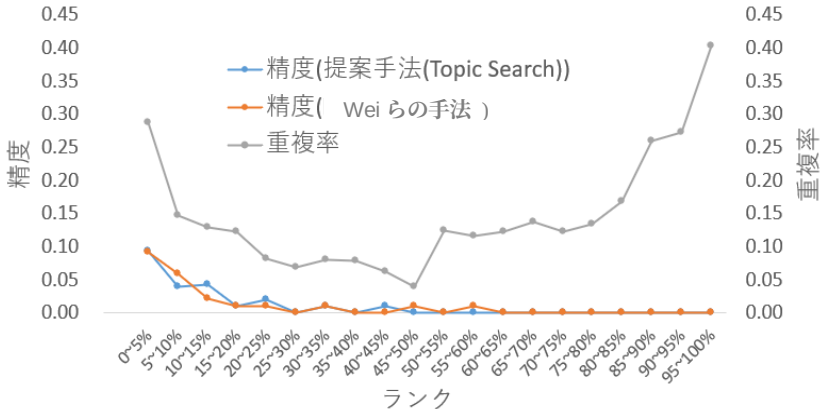

図１　Topic Search と Wei らの検索モデル（ LDA-bsed Document Model ）における精度と重複率の推移

(2) 読むべき論文を特定するための技術開発

抄録にも，研究目的，手法，結果などが記述されており，検索の目的に応じて，読むべき論文を特定するのに役立つ箇所は異なると考えられる．たとえば，手法を提案する研究の新規性を確認する場合や取り組んでいる課題を解決するための手法を探している場合は，抄録中の手法を記述している箇所が読むべき論文か否かの判断に重要であり，従来からの性能向上を達成した研究の新規性の確認では，手法を記述している箇所の他，結果の箇所も重要である．したがって，抄録の各文の役割を推定して，指定された役割の文だけハイライトで示せば，抄録を短時間で確認できると考えられる．また，抄録を確認する際，その分野の英語の専門用語に不慣れな若手の研究者にとっては，専門用語の対訳を示すのも抄録確認の効率を上げるのに有効と考えられる．さらに，読むべき論文をある程度特定した場合，それらと類似する論文を推薦する機能も役立つ．類似文献を探す際，検索者は引用文献等の情報もうまく活用していると考えられる．

さらに，ある特定の領域での研究動向を俯瞰的に見る際，その領域での研究課題の推移，開発された手法や技術の推移などが有効と考えられる．このためには，目的や手法といった抄録中の文の役割の推定および技術を表す用語や手法を表す用語の同定が役立つと考えられる．

そこで，本研究では，読むべき論文を特定する支援システム開発のために，抄録中の各文の役割を推定する手法，日本語と英語の抄録から専門用語の対訳対を抽出する手法，タイトルや抄録に加えてタグ情報や引用文献の情報も活用した類似文献を推薦する手法が重要と考えた．

(3) 検索行動の調査・分析

研究者の検索行動を撮影し，同時にその行為の動機やその行動の結果どう感じたかなど内面の状態を調べるため，検索者に検索中に考えていることを発話してもらいそれを録音する．さらに，検索の前後でのインタビューにより，検索の目的や，検索対象の分野に対する研究歴などを調べる．これらの調査結果を基に，研究者が学術論文検索の際に読むべき論文をどのようにして選択しているのかの傾向を分析する．

上記のような観察調査は，実際にその場で検索を行うことから時間を要し，多くの被験者を対象とした大規模な調査には向かない．そこで，この調査結果を基に，各検索状況に応じて，どのようなデータベースを用いて，どのような情報を基に読むべき論文を絞り込んでいるのかを調査するための質問を作成し，Web 上で２つの大学の工学系の大学院の教員と大学院生を対象に質問紙調査を実施した．

４．研究成果

(1) 情報要求に関連する可能性がある論文を網羅的に収集する検索手法の開発

比較的限定的な情報要求でかつ網羅的な検索結果を要求する検索を対象に，LDA によるトピック分析結果を利用したトピックベースのブーレアン検索手法を開発し，LDA のパラメタの設定が異なる複数のトピックベースのブーレアン検索の結果を統合して検索結果のランキングを行う検索手法 Topic Search を開発した．さらに，Wei らの従来の検索モデルによるランキングモデルと Topic Search によるランキングを統合したランキング手法（Hybrid）を開発した．

NTCIR-1,2 テストコレクションの 40 個の検索課題をデータセットとして用いて，開発したランキング手法である Topic Search および Hybrid と Wei らが開発した従来のクエリ尤度モデルによるランキング手法を評価する実験を行った．各検索手法について，目標再現率を達成するとき，ランキングの上位 何 % を検索結果としなければならないかを求めた．表１に結果を示す．Topic Search と Wei らの検索モデルの比較では，目標再現率が 0.85 以上という高い再現率の場合に Topic Search は Wei らの手法よりランキングの性能が高かった．一方，両者を統合した Hybrid では，目標再現率 0.65～0.95 のすべてで他の手法よりランキングの性能が高かった．たとえば，再現率 0.65 を達成するのに，Wei らのモデルではランキングの上位 18% を検索結果と

表 1 目標再現率を達成する上位 n % の比較

	目標再現率						
	0.650	0.700	0.750	0.800	0.850	0.900	0.950
Topic Search	20%	25%	29%	33%	39%	46%	60%
Hybrid	13%	17%	21%	27%	33%	40%	55%
Wei らの検索モデル	18%	22%	27%	33%	41%	51%	62%

して出力する必要があるのに対し，Hybird では上位 13% を出力すればよい．これは論文を確認する量を約 28% 削減していることになる．

これまで，検索システムの評価は，ランキング上位の文献の内の情報要求に関連する文献の割合が重視されてきた．ランキングのある程度下位まで見た場合の検索漏れが少ないことを指標として検索手法を開発・評価した研究はほとんどなく，本研究はその意味でも意義がある．

(2) 読むべき論文を特定するための技術開発

ニューラルネットワークの１つである LSTM（Long Short-Term Memory）を用いた，論文の抄録中の各文の役割を推定する手法を開発し，Medline から収集した生命科学分野（英語）の構造化抄録に適用して有効性を示した（提案法，SVM ともに precision 約 90%）．しかし，この生命科学分野の構造化抄録で学習した LSTM を情報科学分野の英語抄録（各文に人手で４つの役割を付与）に適用したところ，十分な推定性能は得られず，分野別の文の役割付き抄録コーパスが必要であることが分かった．一方，日本語論文に関しては，構造化抄録を大規模に収集することが困難であったため，情報科学分野の日本語論文 1000 編を対象に，抄録の文の役割を人手で付与したコーパスを構築し，SVM および独自に開発した識別器による抄録の文の役割の推定実験を行った(独自開発の識別器がやや性能が高いが，独自開発の識別器，SVM ともに precision 約 70%)．クロスバリデーションによる推定実験（つまり，識別器の訓練データとテストデータが共に情報科学という同一分野での実験）であったが，Medline から収取した英語構造化抄録を対象とした文の役割推定の性能よりかなり低い性能であった．このことから，1000 編程度の抄録では不十分で，抄録の各文の役割を推定する実用に耐えうる程度の性能のシステムを構築するには，分野ごとにかなり大規模な文の役割付きコーパスを構築することが重要であり，その構築法が課題として残った．

その他，日本語と英語の両方の抄録を入力として日英の専門用語の対訳対を抽出する手法の開発と API として実装，論文のタイトル・抄録・タグ情報・引用文献の情報を活用して類似文献を推薦する深層学習ベースの推薦アルゴリズムの提案と評価実験，技術用語推定技術と文の役割推定技術を用いた指定された分野の研究動向を俯瞰する手法の提案，学術論文集合から学術リソースリポジトリを自動的に構築する手法の提案を行った．

(3) 検索行動の調査・分析

博士課程学生を含む電気・電子・情報工学分野の 10 名の研究者に対して観察調査を行った．これにより，それぞれの，研究領域（研究テーマ），研究歴，読むべき論文かどうかを判断する際に確認する論文の要素（タイトル，抄録，本文，参考文献，出版年，雑誌名等）と確認する順番，利用しているデータベースを調査した．確認する論文の要素とその順番に関しては大きく３つのパターンが観測された．また，データベースから電子ジャーナルへのリンクが一般的となった状況を反映し，論文タイトルを確認後，直接本文の数式を確認する検索者や，Google の画像検索で論文中の図のみを確認して情報要求に関連する論文を見つける検索者など従来と異なる行動が見られた．

上記の観察調査の分析結果を踏まえ，研究段階や検索の目的に応じた，確認する論文の要素や利用するデータベースを調べるための質問紙調査を実施した．被験者は教員 30 名，博士課程学生 18 名，修士課程学生 40 名である．研究分野は観察調査と同じく，電気・電子・情報工学分野である．研究者（教員＋博士課程学生）は，研究を始めた段階／手法を検討する段階／論文にまとめる段階といった研究段階に応じて，良く利用するデータベースが異なっていた．一方，修士課程学生はどの段階でも Google をよく利用していた．また，課題に合った研究手法を見つけるため／詳しくない分野の先行研究を調べるため／最新の研究動向を知るためといった検索の目的に応じて，研究者は，論文の要素（タイトル，抄録，本文，出版年，ページ数，雑誌名，著者名など）や論文本文の要素（Introduction, Related Work, Method, Conclusion, References, 図表や式）が異なっていた．

これらを分析した結果は，研究の段階に応じたデータベースの利用法，検索の目的に応じた論文の確認要素・論文本文の確認要素に関する修士課程学生のような若手の研究者を対象とした図書館職員による講習会等の人的支援に生かせると考えられる．一方，References を一定割合の研究者が確認しており，その意味で，開発した引用文献の情報も活用した類似文献推薦手法は有効と考えられる．また，課題にあった研究手法をみつける際には図表や式を多くの研究者が確認していることから，今後取り組んでいくこととして，図表や式による論文の検索手法や類似論文推薦手法の開発が挙げられる．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計20件（うち招待講演　0件／うち国際学会　16件）

１．発表者名 Satoshi Fukuda, Yoichi Tomiura, Emi Ishita
２．発表標題 Research Paper Search Using a Topic-Based Boolean Query Search and a General Query-Based Ranking Model
３．学会等名 30th International Conference on Database and Expert Systems Applications (DEXA 2019)（国際学会）
４．発表年 2019年

１．発表者名 難波英嗣
２．発表標題 Web上の学術リソースリポジトリの構築
３．学会等名 言語処理学会 第25回年次大会
４．発表年 2019年

１．発表者名 Satoshi Fukuda, Yoichi Tomiura
２．発表標題 A Study for the Support of a Search Formula Creation for the Exhaustive search of an Academic Paper based on a User’s Information Need
３．学会等名 10th Asia Library and Information Research Group (ALIRG) Workshop（国際学会）
４．発表年 2018年

１．発表者名 Satoshi Fukuda, Yoichi Tomiura
２．発表標題 Clustering of Research Papers based on Sentence Roles
３．学会等名 20th International Conference on Asia-Pacific Digital Library (ICADL 2018)（国際学会）
４．発表年 2018年

１．発表者名
Satoshi Fukuda, Yoichi Tomiura

２．発表標題
Toward a Search Formula Creation Support for the Exhaustive Search of an Academic Paper

３．学会等名
Toward Effective Support for Academic Information Search Workshop at ICADL 2018（国際学会）

４．発表年
2018年

１．発表者名
Ho, Xanh, Akiko Aizawa

２．発表標題
Enhancing Collaborative Variational Autoencoder with Tag and Citation Information for Scientific Article Recommendation

３．学会等名
Toward Effective Support for Academic Information Search Workshop at ICADL 2018（国際学会）

４．発表年
2018年

１．発表者名
Hidetsugu Nanba

２．発表標題
Construction of an Academic Resource Repository

３．学会等名
Toward Effective Support for Academic Information Search Workshop at ICADL 2018（国際学会）

４．発表年
2018年

１．発表者名
Emi Ishita, Yasuko Hagiwara, Yoichi Tomiura

２．発表標題
Users' Searching Behavior for Academic Papers

３．学会等名
Toward Effective Support for Academic Information Search Workshop at ICADL 2018（国際学会）

４．発表年
2018年

１．発表者名
Yang Zhao, Zhiyuan Luo, Akiko Aizawa

２．発表標題
A Simple Language Model based Evaluator for Sentence Compression

３．学会等名
56th Annual Meeting of the Association for Computational Linguistics（国際学会）

４．発表年
2018年

１．発表者名
Emi Ishita, Yasuko Hagiwara, Yukiko Watanabe, Yoichi Tomiura

２．発表標題
Which Parts of Search Results do Researchers Check when Selecting Academic Documents?

３．学会等名
18th ACM/IEEE on Joint Conference on Digital Libraries(JCDL'18)（国際学会）

４．発表年
2018年

１．発表者名
Hayato Hashimoto, Kazutoshi Shinoda, Hikaru Yokono, Akiko Aizawa

２．発表標題
Automatic Generation of Review Matrices as Multi-document Summarization of Scientic Papers

３．学会等名
BIRNDL ' 17: Bibliometric-enhanced IR and NLP for Digital Libraries, workshop at SIGIR 2017（国際学会）

４．発表年
2017年

１．発表者名
Yasuko Hagiwara, Emi Ishita, Emiko Mizutani, Kana Fukushima, Yukiko Watanabe, Yoichi Tomiura

２．発表標題
Identifying Key Elements of Search Results for Document Selection in the Digital Age: An Observational Study

３．学会等名
ICADL 2017: 19th International Conference on Asia-Pacific Digital Libraries（国際学会）

４．発表年
2017年

１．発表者名
Satoshi Fukuda, Yoichi Tomiura

２．発表標題
Using Topic Analysis Techniques to Support Comprehensive Research Paper Searches

３．学会等名
IALP 2017: 21th International Conference on Asian Language Processing（国際学会）

４．発表年
2017年

１．発表者名
難波 英嗣

２．発表標題
複数論文概要の解析による特定分野の技術動向分析

３．学会等名
第10回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2018)

４．発表年
2018年

１．発表者名
Yasuko Hagiwara, Emi Ishita, Emiko Mizutani, Yukiko Watanabe, Yoichi Tomiura

２．発表標題
A Preliminary Study and Analysis to Identify Key Elements in Document Selection

３．学会等名
The information behavior conference information seeking in context (ISIC2016)（国際学会）

４．発表年
2016年

１．発表者名
難波英嗣

２．発表標題
レファレンス事例の分析結果に基づいた論文検索システムの構築

３．学会等名
データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2017)

４．発表年
2017年

１．発表者名
Kosuke Furusawa, Hongjun Fan, Yoichi Tomiura, Emi Ishita

２．発表標題
Encompassing Retrieval of Academic Papers for User's Information Need

３．学会等名
17th Asian Digital Library Conference (ICADL 2015)（国際学会）

４．発表年
2015年

１．発表者名
Paul Willot, Kazuhiro Hattori, Akiko Aizawa

２．発表標題
Extracting Structure from Scientific Abstracts

３．学会等名
17th Asian Digital Library Conference (ICADL 2015)（国際学会）

４．発表年
2015年

１．発表者名
Yasuko Hagiwara, Meizhi Wu, Emiko Mizutani, Yukiko Watanabe, Emi Ishita, Yoichi Tomiura

２．発表標題
An Experiment to Identify How Researchers Select Documents from Search Results

３．学会等名
The annual meeting of the Consortium of iSchools Asia-Pacific（国際学会）

４．発表年
2015年

１．発表者名
難波 英嗣

２．発表標題
レファレンス事例の分析による論文検索に効果的な要素の調査

３．学会等名
第8回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2016)

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

ユーザの視点に立った高度な学術論文検索支援に関する総合的研究 http://nlp.inf.kyushu-u.ac.jp/search_assist.html

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	福田　悟志 (Fukuda Satoshi) (10817555)	九州大学・システム情報科学研究院・助教 (17102)	
研究分担者	石田　栄美 (Emi Ishita) (50364815)	九州大学・附属図書館・准教授 (17102)	
研究分担者	相澤　彰子 (Aizawa Akiko) (90222447)	国立情報学研究所・大学共同利用機関等の部局等・教授 (62615)	
研究分担者	難波　英嗣 (Nanba Hidetsugu) (50345378)	広島市立大学・情報科学研究科・准教授 (25403)	