# 令和5年度修士論文題目一覧

統合新領域学府ライブラリーサイエンス専攻修士課程

| 学位 | 学生氏名 | 研究課題 | 論文公開可否 |
|---|---|---|---|
| 修士<br>（ライブラリーサイエンス） | 佐々木　陸 | 読むべき最新論文の判別方法 | 可 |
| 修士<br>（ライブラリーサイエンス） | 篠原　桐 | 流通を考慮した研究データのメタデータ付与支援 | 可 |
| 修士<br>（ライブラリーサイエンス） | 山﨑　基数 | 研究内容からの研究データ検索 | 可 |

# 読むべき最新論文の判別方法

A Method for Identifying Recent Papers to Read

2FS22201G　佐々木 陸　SASAKI Riku

新たな研究着手の際、その研究テーマに関して，扱っている課題，手法とその性能，理論などの周辺トピックを調査する．その際，レビュー論文に引用されている論文を調査することは一つの有用な手段である．しかし，実際ある研究テーマで用いられている手法やその性能などに関して、最新の動向を論じている論文は，まだレビュー論文によって引用されていない場合がほとんどである．引用データベースを用いて，レビュー論文で紹介されていた特定の論文を引用している比較的新しい論文を見つけることは可能だが，引用の理由はさまざまであり，引用元論文すべてが同じトピックを扱っているとは限らない．また論文の質を問わず，多数の論文が見つかるうえに，被引用数といった客観的な評価にもまだ差がないため，読むべきものを絞るのが難しいという問題点もある．そのため，最新の論文のサーベイは一般的な論文のものと比較して，より手間と時間を要する作業となっている．

そこで本研究では被引用論文の抄録，引用元論文の抄録，導入，結論，引用文に基づいて，引用関係にある２つ論文において，引用元論文が被引用論文と同じトピックを扱っており(条件①)，かつそのトピックを代表するよい(条件②)論文を判定する方法を提案する．ここでいうトピックとは研究対象やそれに対する主要なアプローチを指す．この方法を用いて，与えられた研究トピックの最新の論文を見つけることができる．まず，レビュー論文等を用いてその研究トピックにおける代表的な論文を１つ見つけ，これをキー論文とする．引用データベースを用いて，キー論文を引用している論文を見つけ，提案法に基づいてキー論文を引用している論文がキー論文と同じトピックを扱っており，かつそのトピックを代表するよい論文かどうかを判定する．これによってキー論文を引用している論文の中で読むべき論文を厳選することができる．これによって自身の研究トピックを扱うレビュー論文で引用されている論文や先行研究と引用関係にある論文を調べる時間を大幅に削減することにつながる．また，それらの中でまたキー論文を設定し，その引用論文を判別することで上記の ２つの条件を満たすより新しい論文を見つけることができ，最終的には客観的評価がまだない最新の論文においても先述した２条件を満たすものを判別することができる．

本研究では引用関係にある論文ペアのうち，先述した２つの条件をともに満たすものの判別を自作したデータセットで学習した判別モデルを用いて機械的に行い，最終的には判定結果を Recall，Precision で評価する．情報検索の一種で，見逃しが少ないことを評価するために Recall を重視する．

判別モデルでは被引用論文の抄録，引用元論文の抄録，導入，結論，引用文を入力とし、自然言語モデルである BERT を用いて，入力テキストの分散表現を求め，テキストを数値として扱えるようにする．最終的な出力は入力が先述した２つの条件を満たす正例である確率を表し，閾値 0.5 で正例・負例に判別する．

データセットは，データ作成者（annotator）の主観の影響がでないように，機械的な方法でデータを作成する．レビュー論文はある特定の研究トピックを中心に論文をサーベイしたものであるから，同じレビュー論文で引用（紹介）されている論文は，同一のトピックを扱っていて，そのトピックにおいて代表的な論文であると考えられる．レビュー論文で引用（紹介）されている論文 X と X を引用している論文 Y のペアでデータセットを作成した．Y がレビュー論文で引用されているものを正例，引用されていないものを負例とした．

本実験は Google Colaboratory 上で Pytorch を用いて行った．学習率やモデルの層の出力サイズなど，様々なパラメータを調節して実験を行い，評価には 5-Fold クロスバリデーションを用いた．データセットを５つに分け，モデルの学習とテストをそれぞれ異なる部分で行う．そして、5 回の結果の平均を評価した．その結果，最終的に Precision は約 83%， Recall は約 87%を達成した．

また，データセットの引用元論文の被引用数に正例と負例の間で差があったことから， 追加検証において判別モデルの判定結果と引用元論文の被引用数の関係を調べた．その結果，この判定モデルは基本的には論文の質だけで判別を行っているわけではないが，ある一定以上の被引用数を持つ論文については，その質のよさのから正例と判定されやすいことがわかった．

本研究では引用関係にある２つの論文間の判定を行い，最終的に Precision は約 83%，Recall は約 87%という高い分類精度を達成したが，今回は比較実験を行わなかったパラメータも複数存在し，それらを含めてより細かく調整を行うことで，同モデルでもさらなる精度向上が見込まれる．この判定方法を用いて，１ つのキー論文から条件①，②をともに満たす最新の論文を検索できるシステムができると，より使いやすいものになる考ことが期待される．また本研究で作成したデータセットの論文は本研究トピックである引用分類に関するものを中心に集めたが，他分野の論文においても同様の精度を達成できれば，より多くの研究者の役に立てるものになると考えている．

# 流通を考慮した研究データのメタデータ付与支援

Support for adding metadata to research data in consideration of distribution

2FS22202R　篠原 桐　SHINOHARA Kiri

研究データとは，研究のために収集・作成・観測したデータで，研究の成果である論文の根拠となるデータ，及び研究成果そのものであるデータの両方を含むものである．近年，オープンサイエンスの高まりにより，研究データの公開が促進されており，その利活用によるイノベーション創出が期待されている．Scopus に掲載されているデータ論文の掲載数も，2014 年では 2 件だったのに対し，2023 年では 7341 件とその数を大きく伸ばしている．また，Web of Science の Data Citation Index によると，毎年 100 万件近くものデータセットが登録されている．公開されている研究データを利用するためには，利用者が研究データを検索し，それを発見できる必要がある．

研究データを検索する際は，一般的に，研究データに付与されているメタデータを通して検索が行われる．研究データが広く利用されるためには，研究データが作成された分野とは専門の異なる利用者からも検索される必要があるが，専門としていない分野の検索クエリを思いつくことは困難である．そのため，研究データの流通性を高めるためには，メタデータを付与する際に，専門としていない利用者が検索しやすいメタデータを付与する必要がある．

一般的に，研究データに付与するメタデータの登録は，その研究データの作成者が行う．しかし，作成者によるメタデータだけでは，記述が不十分である場合も多く，有効な検索が行えない可能性がある．

そこで，本研究では生成 AI を利用することで研究データの作成者が研究データにメタデータを付与する際に，研究データが検索される可能性の高い，流通することを考慮したメタデータの付与の支援ができるのではないかと考え，その可能性について調査を行った．

本研究では，生成 AI のモデルとして ChatGPT を利用した．Scopus からデータ論文を収集し，各データ論文に対し，ChatGPT に研究データのタイトルと，データ論文に書かれた概要を入力し，このデータを利用する可能性のある研究を考慮したキーワードを，与えた情報を含まないよう 10 個提案させた．提案されたキーワードを調査するために，主観評価と客観評価の２種類の評価を行った．主観評価では，48 個のデータ論文に対し，提案されたキーワードを５段階で評価した．評価は数字が大きいほど高いものとなる．客観評価では，50 個のデータ論文に対し，ChatGPT に提案されたキーワードを検索クエリとし，Semantic Scholar を利用して，そのデータ論文を引用した論文が検索可能かを調査することで，間接的に評価を行なった．

主観評価に用いた 48 個のデータ論文に対し提案された合計 480 個のキーワードの中で，与えた情報に含まれていない新たなキーワードは実際には 401 個だった．データ論文あたり平均 8.35 個の新たなキーワードが提案された．

提案されたキーワードの主観評価では，自身で評価を行うと共に，その評価が信頼できるかを確認するために，第三者に同様に主観評価をしてもらい，その結果との $\kappa$ 係数を調べた．第三者による評価では，48 個の研究データのうち，10 個は評価基準の学習に利用し，残りの 38 個の研究データに対して評価を行った．その結果，自身の評価の平均スコアは 3.45，第三者による評価の平均は 3.24 となった．$\kappa$ 係数は，0.50 となり，十分な一致を示すことはできなかった．評価 3 以上のキーワードは，研究データにメタデータとして付与できると考えられるため，評価を 3 以上と 2 以下の 2 値で分類し，その 2 値に対して $\kappa$ 係数を調べたところ，0.66 となり，十分な一致を示すことができた．一致度が高く，評価 3 以上のものが多いことから，提案されたキーワードは，メタデータとして付与される可能性のあるものが多く提案されていると考えられる．

間接的な客観評価では，50 個のデータ論文に対し，31 個のデータ論文において，そのデータ論文を引用している論文を検索することができた．また，検索に成功した 31 個のデータ論文に対し，元々持っていたキーワードを検索クエリとして，同様にそのデータ論文を引用している論文を検索したところ，18 個のデータ論文では，元々持っていたキーワードでは検索を行うことができなかった．このことから，この 18 件において，提案されたキーワードをメタデータとして付与することで検索される可能性が高まると考えられる．

新たなキーワードの提案数，主観評価によるキーワードの有用性，客観評価によるキーワードの妥当性から，メタデータの付与の支援に有効に利用できる可能性が高いと考えられる．

本研究では，生成 AI を利用することで，研究データのメタデータとして，利用用途を考慮したキーワードの提案を支援できると考え，その可能性について議論した．その結果，新たなキーワードの候補の推薦に有効である可能性が示された．しかし，検索結果の向上については，利用法を考慮したキーワードを付与できる可能性を示すことはできたが，直接的な客観評価をすることができないことなどから，可能性を示唆するに留まった．今後は，主観評価を行う人数を増やす，評価を行うデータ数を増やす，データの偏りをなくすなどを行うことで，提案されたキーワードの有効性についてより議論していく必要がある．

# 研究内容からの研究データ検索

Research data retrieval by using research summary

2FS22203N　山﨑 基数　YAMASAKI Motokazu

近年,インターネットをはじめとする情報技術の発達により急速なペースで蓄積される研究データは,科学研究や技術の進展において不可欠なものとなっている.新しい発見や知見を得るためには,これらのデータを適切に利用して研究を行うことが重要である.研究データの検索はデータに付与されたメタデータに対してのキーワード検索によって行われる．しかし，従来のキーワードベースの検索手法ではユーザが適切なキーワードやフレーズを入力する必要があり,目的の研究データに関する詳細な情報を持っていない場合は十分な検索結果を得ることが難しい.そこで，本研究では研究データ検索における LLM（大規模言語モデル）の有効性を検討した．LLM に基づいた ChatGPT を活用し,より研究者が欲しい研究データにたどりつきやすくするための手法を提案する．ChatGPT は LLM に基づいているため,ユーザが入力した文章に対して人間のように自然な回答を行うことができる.ユーザが ChatGPT に自身の研究の概要を伝えることにより関連する研究データや役に立つ研究データについて情報を得られる可能性があると考えた．キーワード検索と異なり，LLM を利用することで入力した文章から関連のある研究データを推測し,ユーザが特定の研究データに関する詳細な情報を持たない場合でも目的の研究データを得やすくなる可能性があると考えた.

本研究では研究者が計画している研究の内容を文章で表現し,その文章から自身の研究を行うのに役に立つ研究データを入手するのに ChatGPT が有用であるか検証した.しかし,実際に研究者に研究内容を入力してもらい実験を行うことは多くの研究者の協力を得る必要があることから困難である．そのため，我々は ChatGPT に既存の論文の Abstract を入力し，その論文に書かれている研究の役に立つ研究データを提案させることを考えた.

本研究は実験Ⅰを行った後に,実験Ⅰで得た知見をもとに実験Ⅱを行った．基本的な実験の手順は同じである．実験 Ⅰ と実験Ⅱの相違点としては主に使用した論文や研究データの変更である．まず、実験Ⅰと実験Ⅱで共通する手順について説明する．ChatGPT はある時点までの情報を学習しており,それに基づいて回答を生成している.ChatGPT が学習にしていない論文の Abstract を利用することで研究者がこれから行う研究の概要を入力することを再現した．具体的には既存の論文の Abstract と研究データを ChatGPT に提案させるための文章を入力した．研究データ,正確には研究データに付随するデータ論文を 50 個集めた.また,研究データ 1 つにつき研究データを引用している論文を 2 本集めた．つまり研究データ（に付随する論文）50 個とそれを引用している論文 100 本に対して実験を行った．ChatGPT に提案させた研究データの中に目的の研究データ(研究データを引用している論文とペアとなっている研究データ）が含まれているかを調査し，含まれている場合には「発見」，含まれていない場合には「未発見」としてその数をカウントした.

実験Ⅰでは 100 本の論文のうち 65 本で研究データの発見をすることができた．しかし，実験Ⅰで使用した論文には Abstract に研究データ名が直接記載されているものや研究データを引用しているが研究の本筋と研究データが結びついていないものが含まれていた． Abstract に欲しいデータの名前が記載されている論文は研究者が目的の研究データに詳しくない状況を再現するのに不適切であると考えた．また，研究データが論文内で活用されていないものも研究者が自身の行いたい研究の役に立つデータを探す状況の再現を行う上で不適切な論文であると考えた．そのため， Abstract に研究データ名が記載されている論文,研究データが活用されていない論文を差し替えて実験Ⅱを行った．実験Ⅱでは，100 本の論文のうち 67 本の論文で研究データを発見できた．更に，キーワード検索との比較を行うために実験で使用した論文の Abstract から検索ワードを抽出し，それを用いて Google Dataset Search と Google Scholar のそれぞれで検索を行った．キーワードの選定に公平性を持たせるために ChatGPT を用いて Abstract からキーワードを抽出した.Google Dataset Search，Google Scholar での検索結果上位 30 件における研究データ発見数はそれぞれ 17 個と 23 個で ChatGPT に提案をさせた場合と比べて少なかった.

本研究では研究データ検索における LLM の利用可能性を調査するために ChatGPT を用いて研究内容からの研究データ検索が行えるか実験した．実験の結果，実験で使用した 100 本の論文のうち，67 本の論文において目的の研究データを発見することができた.これにより LLM を利用することが研究データの検索を行うのに役立つ可能性を確認した．今後の課題としては，目的のデータによりたどり着きやすくするための質問文の工夫を行うことや回答内容について更なる分析を行うことが挙げられる．また，見つかった研究データや ChatGPT が頻繁に提案した研究データのメタデータを分析することで研究データを見つけやすくするためのメタデータについて知見を得られる可能性がある.このことは研究データを従来のキーワード検索で見つけやすくすることにも寄与する可能性がある.