令和７年度修理論文題目一覧

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	氏　名	論文題目	論文公開可否
修士 （ライブラリーサイエンス）	久保　俊介	固定カメラによる動画を活用した効率的な資料デジタル化手法の開発	可
修士 （ライブラリーサイエンス）	林田 慎太郎	生成AIを用いた論文からの糖類の相対反応性情報の抽出	否
修士 （ライブラリーサイエンス）	劉　子衿	研究データ発見性向上に向けた生成AIによるキーワード推薦とその評価	否
修士 （ライブラリーサイエンス）	徐　雨戈	台北帝国大学関係文書の研究	可
修士 （ライブラリーサイエンス）	山仲　一颯	解答コードの確率的表現に基づく類似性観点可変なプログラミング問題検索	可
修士 （ライブラリーサイエンス）	黒川 怜雄	データセット推薦RAGのための関連論文ランキング手法	可

固定カメラによる動画を活用した効率的な資料デジタル化手法の開発

Development of an Efficient Document Digitization Method Based on Video Captured by a Fixed Camera

2FS24201P　久保　俊介　KUBO Shunsuke

資料のデジタル化は，知識資源の保存と活用の観点から極めて重要である．デジタル化により，劣化しやすい原本への物理的接触を減らしつつ，インターネットを通じた広範な公開や OCR による全文検索が可能となる．このような背景から所蔵資料のデジタル化は推進されているが，予算や人員が限られる小規模機関においては費用や作業効率の面で課題が伴う．専用のドキュメントスキャナーは導入・維持費用が高額であり，費用面での負担が大きい．デジタルカメラを用いた手動撮影は安価な代替手段となるが，ページごとに「固定・確認・撮影」のプロセスを繰り返す必要があり，作業効率が低いという課題がある．

これに対し，ページをめくる様子を動画として撮影し，静止したページ画像をソフトウェアで自動抽出する「動画からのドキュメントキャプチャ」が注目されている．しかし，既存の手法は，ページめくりの速度変化や照明条件の変動に弱く，また，事前に文書データで学習させたモデルを必要とするなど，汎用性や学習コストに課題があった．本研究では，これらの課題を解決し，一般的なデジタルカメラとノート PC のみで動作する，追加の学習を必要としない，ページめくりの速度変化や照明変化に頑健なドキュメントデジタル化ワークフローを提案する．

本研究のワークフローは，特徴抽出，2 段階フィルタリング（異常検知，クラスタリング），代表フレーム選択，後処理（手検出，重複検出）という流れで構成されている．特徴抽出では，入力動画の各フレームから，ImageNet で事前学習済みの軽量 CNN の中間層を用いて特徴量を抽出する．これにより，高性能な計算資源を搭載していない一般的なノート PC でも実用的な速度で，特定の文書に対する追加学習を行わずに，画像を視覚的に捉えることが可能となる.2 段階フィルタリングでは，ページめくりシーンを排除し，静止したページのフレームを残すことを目的とする．第 1 段階の異常検知では，連続するフレーム間の特徴量ベクトルのコサイン類似度を算出し，類似度が閾値を下回る場合，視覚的に大きな変化があったとみなし，これを「異常」として除外する．これにより，ブレや変形を含む不適切なフレームを大まかにフィルタリングする．第 2 段階では，異常検知を通過した「正常」フレームに対し，密度ベースのクラスタリング手法を適用する．これにより，同じページを映している類似したフレーム群をクラスタとしてまとめると同時に，クラスタを構成する最小枚数に満たないフレーム群をノイズクラスタとして除去する．密度ベースの手法ではクラスタ数を事前に指定する必要がないため，ページ数が未知の資料にも適用可能である．既存手法の RGB 変化を利用するものに比べ,CNN 特徴量での類似度変化は照明変化に強く，クラスタリングは似たフレームを同じクラスタに割り当てるものなので，ページめくりの時間変化に影響されない．クラスタリング後，各クラスタの時系列的な中央フレームを代表画像として選択し，手が大きく映り込んでいる画像を除外する．さらに，抽出された画像間の類似度を計算し，フレーム番号の距離を考慮した上で重複ページを削除する．

提案手法の有効性を検証するため，異なる条件下で撮影された 4 つの独自データセットおよび公開データセットを用いた評価実験が行った．独自データセットでの評価において，ページめくりの速度や安定性が異なる 4 種類の動画すべてにおいて Recall は 1.0 を達成した．これは，デジタルアーカイブにおいて致命的となる「ページの欠落」が一切発生しなかったことを意味する.「余分なページ」を意味する Precision は 0.67 から 0.91 と変動したが，これはガラス棒の使用や焦点のずれなど，ページめくり以外の動作に起因するものであり，余分な画像は後処理で容易に削除可能であるため許容できる．公開データセットである PUCIT Page Turn dataset での比較において，既存手法との比較で提案手法は Recall が 0.99，F1 スコアが 0.96 という極めて高い性能を示した．特に，教師あり学習を必要とする手法を Recall において上回る結果となり，事前学習済みモデルのみを用いた本手法の実用性の高さが実証された．

本研究は，高価な機材や高性能な計算資源，個別の学習データを必要とせず，固定カメラとノート PC のみで，既存手法を上回る精度のドキュメントキャプチャを実現した．異常検知とクラスタリングを組み合わせることで，ページの欠落の少ないページめくりシーンの除去が行えることを確認した．今後の課題として，部分的な指の検出精度向上，ピントのずれた画像のフィルタリング，さらに高品質な画像抽出のために，同一ページを写すフレーム複数枚を統合する超解像技術の導入などが挙げられる．本手法は，費用面と作業時間のリソースが限られた小規模な図書館やアーカイブ機関において，資料のデジタル保存を加速させるための現実的かつ有効な解決策であると結論付けられる．

台北帝国大学関係文書の研究

A Study on the Documents of Taihoku Imperial University

2FS24204G　徐 雨戈　XU YUGE

台北帝国大学は日本統治下の台湾に設置された帝国大学であり、台湾における最初の高等教育機関であった。1945 年の日本の敗戦後、中華民国政府により接収され、現在の国立台湾大学の前身となった。台北帝国大学に関する研究は、設立目的、学部構成、研究活動、戦後期における日本人教員の留学政策に至るまで多岐に蓄積されてきたが、台北帝国大学が作成した公文書群そのものを対象とする研究は極めて乏しく、特に現存する台北帝国大学関係文書の規模・性格や、接収に伴う文書移管の問題については、体系的な検討が進んでいない。

本研究は、現存する台北帝国大学関係文書の全体像を明らかにすることにより、台北帝国大学研究および台湾植民地研究の基盤を構築すること、大学組織の改編に伴う公文書移管の問題を検討することを目的とする。

現存文書の調査では、日本および台湾の主要機関――国立公文書館、国史館台湾文献館、台湾中央研究院近代史研究所、および国立台湾大学附属図書館に所蔵される台北帝国大学関係文書を中心に調査し、分析を行った。

国立公文書館に所蔵される台北帝国大学関係文書は 650 件であり、作成年は 1928 年の大学設立期から大学接収後の 1948 年まで広がる。作成部局は内閣関係機関が 640 件を占め、台北帝国大学自身による作成文書は極めて少ない。内容別では人事関係文書が 509 件と大半を占め、制度・官制改正関係の文書が 133 件、その他は少数にとどまる。

台湾総督府文書については計 4,104 件の文書が確認された。作成年は台北帝国大学設立期から大学移交後に及び、台湾総督府官報掲載文書が 3,614 件と全体の約九割を占める。内容分類では人事関係文書が 2,416 件と過半を占め、制度改正関係や大学運営関係文書も一定の数含まれる。

台湾中央研究院近代史研究所档案館には、台北帝国大学附属医院精神科の患者記録 7,815 件が所蔵されている。1930 年から 1954 年にかけて作成されており、台北帝国大学期と戦後初期の国立台湾大学期を連続的に収録している点が特徴である。ただし、閲覧は館内に限定され、利用には審査を要するなど、アクセス上の制約が存在する。

台湾大学図書館における台北帝国大学関係資料の所在確認の結果、校史資料に文政学部および南方人文研究所を中心とする簿冊計 47 冊が所蔵されていることが判明した。しかしその数量は総体としてごく少なく、内容も学生資料・人事関係文書等に限定される傾向がある。

以上の調査結果から、現存する台北帝国大学関係文書は日本および台湾の複数機関に分散して保存されていることが明らかとなった。ただし文書の作成主体に着目すると、現存文書群の大半は内閣や台湾総督府による制度公布、人事発令等の行政処理に関わるものであり、大学内部の教育活動や業務過程を示す文書は極めて少ない。

続いて、本研究ではその現状が生じた要因を探るため、台北帝国大学の文書管理制度および台湾大学への移交経緯を検討した。台北帝国大学規程の「臺北帝國大學文書取扱規程」および「臺北帝國大學文書ノ編纂及保存ニ關スル規程」に基づき、文書の受領・作成・決裁・保存・閲覧・廃棄に至る制度的枠組みを分析した。

さらに現在台湾大学档案館に所蔵されている大学移交の際に作成された「台北帝国大学移交清冊」に記載された移交公文書目録を取得し、分析を行った。公文書目録を部局ごとに分析した結果、移交公文書目録は人事・会計・例規等の文書および学生関係文書が中心であった一方、議事録等の大学内部の重要な運営事項を示す文書はほとんど確認されなかった。また、一部部局において、存在が想定される文書が欠落していることが確認された。加えて、移交経緯の調査から「移交清冊」の作成期間が短く、作成基準も部局ごとに異なる可能性が高く、実際に移交された文書は目録と完全には一致していなかった可能性があると考えられる。

また、現存文書と「移交清冊」の公文書目録の比較から、現在台湾大学に所蔵される台北帝国大学関係文書は移交文書の一部に限られ、大学移交後の段階で文書保存および移管が体系的に実施されなかったことに起因する文書の減失・散逸が発生した可能性があると考えられる。

戦後の国立台湾大学において旧台北帝国大学文書がいかなる方針のもとで引き継がれ、評価され、保存・廃棄されてきたのかは、具体的過程がなお十分に解明されていない。今後の課題として、戦後初期における台湾大学の移管文書取り扱いの方針の検討と、他の関連研究機関に分散する可能性がある文書群の更なる調査が求められる。

解答コードの確率的表現に基づく類似性観点可変なプログラミング問題検索

Programming Problem Search with Adjustable Similarity Perspectives Based on Probabilistic Representation of Solution Codes

2FS24205R　山仲　一颯　YAMANAKA Issa

近年，データサイエンスや人工知能技術の発展に伴い，これらを支える基盤としてのプログラミング教育の重要性が一層高まっている．初等教育段階からプログラミングが導入されるなど，論理的思考力や問題解決能力の育成を目的とした教育的取り組みが広がる中，学習者の進捗データを活用したパーソナライズされた学習支援の実現が求められている．特に，学習者の理解度に応じて適切な課題を提示する問題推薦は，学習効果を高める上で有効な手法として注目されている．問題推薦には，ユーザ履歴に基づく協調フィルタリング，問題文やメタデータに基づくコンテンツベースフィルタリング，あるいはそれらのハイブリッドな手法が存在する．これらはユーザ間や問題文間，メタデータ間の類似性を基に問題推薦を行う．しかし，協調フィルタリングは学習者の嗜好を反映しやすく，知識獲得を重視する教育目的には必ずしも適さない．また，コンテンツベースフィルタリングにおいては，問題文に基づく推薦では，解答に必要とされるプログラミング知識や構文要素を十分に捉えられない．メタデータに基づく推薦ではアノテーションコストや主観性の問題が存在する．これに対し本研究では，解答コードに基づいて問題の特徴を捉える手法に着目する．コードは問題解答に必要な知識や構造を直接的に反映するため，より客観的かつ，本質的な類似性評価が可能である．しかし，コードベースの類似検索の既存研究でも，類似性の観点が固定的なものとなっているという課題が存在している．プログラミング問題における類似性は本質的に一意に定義できるものではなく，目的や用途によって大きく変化する．例えば，字句的な構造の類似，使用されるアルゴリズムの共通性，文法構造の類似，難易度など，複数の観点が存在する．しかし既存研究では，これらの観点を柔軟に切り替える仕組みが十分に整備されていない．本研究は，解答コードの文法要素に着目し，目的に応じて類似性の観点を変更可能な類似問題検索手法を提案する．

提案手法では，解答コードから抽象構文木(AST)を生成し，その親子関係を生成規則とみなして疑似的な文脈自由文法(CFG)を構築する．さらに，同一問題に対する複数の学生の解答コードを集約し，各生成規則の出現確率を計算することで，疑似的な確率的文脈自由文法(PCFG)として問題を表現する．この PCFG は，特定の文法要素（例：FunctionDef，For，While，If）における，生成規則の確率分布を表している．このアプローチの利点は，単一の解答コードに依存せず，学習者集団が示す多様な解答パターンを確率分布として包括的に表現できる点にある．本研究では，問題間の類似度を，着目する文法要素における生成規則の確率分布間の Jensen-Shannon Divergence (JSD) として定義する．これにより，「for 文の構造が似ている問題」「if 文内の処理が似ている問題」「関数全体の構造が似ている問題」といったように，目的に応じた類似問題検索が可能となる．

実験では，大学初年次向け Python 授業の 47 問の問題と 130 名の学生の解答を用いて類似問題検索を行った．文法要素ごとに類似度行列を可視化した結果，着目する文法要素に応じて，類似問題の傾向が異なっており，その具体的な検索結果も妥当な結果であった．また，FunctionDef に基づく類似度は，CodeBERT や GraphCodeBERT,TF-IDF といったコード全体を捉える既存のコードベースの手法と同様の類似問題の傾向を示し，For や While に基づく類似度は既存手法と異なる傾向を示した．すなわち，提案手法は着目する文法要素を変更することで，検索目的に応じた類似度検索が可能な手法である．また，教員による類似問題検索の人手での性能評価では,FunctionDef に基づく検索は既存手法と遜色ない精度であった．全問題における平均としては，For や If に基づく検索は既存手法には劣るが，既存手法とは異なる類似問題を検索できていた．また，総合的な知識が問われる問題においては，既存手法や FunctionDef よりも高い精度となっている．以上より，提案手法は FunctionDef に基づく検索により既存手法の代替が可能であり，さらに，他の文法要素に基づく検索によって，目的に応じて観点を変更可能な手法であるため，類似問題検索において有効である．

一方で本研究にはいくつかの限界が存在している．実験に用いたデータセットは 47 問と小規模であるため，より大規模で多様なデータセットでの検証が課題である．また，AST の親子関係を基に CFG を構築するため，孫要素以降の深い構造的特徴や，コードの逐次的な順序，変数名，リテラル値といった詳細情報が失われている．本研究では類似問題検索時に 1 つの文法要素のみに着目している．そのため，複数の文法要素に基づく検索といった複合的な類似性の観点による類似問題検索が今後の展望である．

データセット推薦 RAG のための関連論文ランキング手法

A Related-Paper Ranking Method for Dataset-Recommendation RAG

2FS24206N 黒川 怜雄 KUROKAWA Reo

近年、オープンサイエンスの潮流に伴い、公開されるデータセットの数は爆発的に増加している。これによりデータ駆動型科学の加速が期待される。しかし、現状の検索システムの多くは、データセット作成者が付与したメタデータ(タイトルや説明文、キーワード)を検索対象としているため、記述の欠損や入力語彙との不一致といった要因により、研究者の研究目的に適したデータセットを検索結果から発見するのが困難になってきている。これに対し、研究概要を入力し、大規模言語モデル(LLM)の高度な言語処理能力を活用して、研究目的に適したデータセットを推薦することが期待できるが、LLM のみを利用した推薦手法では、LLM の事前学習に含まれない情報は扱えないため、事前学習時点より後に公開されたデータセットを推薦することはできないという問題がある。さらに、ハルシネーションというリスクがある。

そこで本研究では、外部知識を検索によって補いながら生成を行う Retrieval-Augmented Generation (RAG) に着目し、その推薦品質を決定づける「検索器」の高度化を目的とする。具体的には、研究概要を入力として、関連論文を上位に提示する「関連論文ランキング手法」を提案する。関連論文とは、入力が表す研究で使われる可能性のあるデータセットを利用している論文である。この手法により、研究テーマや手法が類似した先行研究のデータ利用実績に基づいた、信頼性の高いデータセット推薦が可能になると考えられる。

提案手法では、計算コストと精度の両立を図るため、二段階ランキングモデルを採用した。第 1 段階(Ranker)では、入力文と候補論文を独立に符号化するモデルを用いて、論文データベースから候補を高速に回収する。第 2 段階 (Reranker) では、入力と候補の相互作用を考慮したモデルを用いて、第 1 段階で得られた上位候補 (30,000 件) に対して精密なスコアリングを行い、最終的な順位を決定する。

本研究では、実際に研究者に研究概要を入力してもらい、上位にランキングされる論文が研究目的に適したデータセットを使っているかを評価してもらう代わりに、約 1,160 万件規模の科学論文のメタデータを含む大規模コーパス S2ORC(Semantic Scholar Open Research Corpus)を用いた評価実験を設計した。具体的には、S2ORC 中のランダムに選択したデータ論文(50 件)のデータセットを利用している S2ORC 中の論文をランダムに 1 つ選択し、このアブストラクトを提案手法の入力 q とする評価実験を行った。入力 q と同一のデータセットを利用している S2ORC 中の残りの論文アブストラクトを R(q)とし、提案手法により、S2ORC の全論文アブストラクトをランキングし、その上位に R(q)の要素のアブストラクトが出現するかで評価した。当該のデータ論文のデータセットを利用しているか否かは、そのデータ論文を引用しているかを機械的に判定し、引用しているものについては LLM による判定を行い、評価用データに関しては候補論文のすべての論文を目視で確認した。

本タスクに適したモデル構成を明らかにするため、Ranker と Reranker それぞれについて、複数の事前学習済みモデルや学習条件（損失関数や追加学習手法）を試行し、最良のものを選び評価を行った。

実験の結果、Ranker では上位 30,000 件までに正例（入力アブストラクト q に対する R(q)の要素）を含む割合は 0.80 であり、この上位 30,000 件に対して Reranker では上位 300 件に正例を含む割合は 0.46 であった。

Ranker により約 1,160 万件をランキングし、その上位 30,000 件に正例が 1 つでも含まれる割合が 0.8 であることから、アブストラクトの情報だけからも同一データセット利用をある程度識別できると言える。また、Ranker によるランキングの上位 300 件に正例が含まれる割合は 0.36 であり、Reranker では上位 300 件に正例が含まれる割合が 0.46 であったことから、二段階ランキングの有効性が示せた。0.46 という割合は実用を考慮するとまだ十分な値とは言えない。データセット推薦システムとしては、ランキング上位のアブストラクトに対応する論文テキスト、研究概要、および指示を LLM に入力することを想定している。そこで今回の実験では、LLM のコンテキスト長と論文テキストサイズを考慮し、入力できる論文の数を 300 とした。RAG におけるコンテキスト圧縮技術等を利用すれば、10〜20 倍の論文を LLM への入力に埋め込むことができる。Reranker によるランキングの上位 900 件では 0.60、上位 5,000 件では 0.70 に達しており、実用の可能性も見えてくる。

残された課題としては、評価用の入力（論文アブストラクト）を増やし、信頼性の高い評価実験を行うこと、本方式で絞り込んだ関連論文を RAG として組み込んだ LLM によるデータセット推薦の性能の評価が挙げられる。