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　 ５ 月 ２９ 日現在

機関番号：３２６６５
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２５８５

研究課題名（和文）目次の分散表現による図書の概念検索手法の研究

研究課題名（英文）Study on a Concept Extraction Method for Book Search using Embedding Model of Table of Contents

研究代表者
関　亜紀子（SEKI, Akiko）

日本大学・生産工学部・講師

研究者番号：６０３８６６７０

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：本研究では、大学などの学校図書館での図書探索を対象とした対話形式での図書の推薦手法について検討している。ここでは、対話文に含まれる曖昧な表現から、検索に有用な専門用語を類推し、検索クエリー用の語彙を拡張するための特徴語の抽出手法の検討を行っている。図書の内容としての類似度を計算できるようにするために、目次などの書誌情報を用いて専門書の分散表現モデルを構築し、既存のモデルとの比較検証を行った。また、プロトタイプを構築し、対話による検索クエリーの拡張と類似図書の綴り込みの実現手法を提案した。

研究成果の学術的意義や社会的意義
本研究の学術的意義や社会的意義は、図書の目次情報と大規模言語モデルを用いて、図書の分散表現モデルを構築することで、従来のキーワード検索では得られない類似図書の探索が可能になることを示した点にある。ここでは、対話から得られた曖昧なキーワードから、利用者を支援するための検索クエリーの拡張手法を提案している。これにより、利用者とのインタラクションの中で図書探索の目的やテーマを具体化させることが可能であり、利用者による図書探索の支援だけでなく、学校図書館の司書の業務支援などへの応用も期待できると考えている。

研究成果の概要（英文）：In this study, we focused on supporting book searches in academic libraries, which is investigated a book recommendation method using an interactive system. Automatic keyphrase extraction and expand queries from dialogue sentences have become an important task in this system. So, we proposed a method to expand queries to include technical terms based on ambiguous expressions contained in dialogue sentences. To calculate the similarity of book content, we also constructed a sentence embedding model of specialized books based on the table of contents. And, we have developed a prototype to realize book recommendation through dialogue.

研究分野：情報工学

キーワード：図書検索　自然言語処理　分散表現

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

従来、図書検索では、タイトルを用いた件名検索や、著者や出版などの書誌情報を用いた書誌検索、日本十進分類(NDC)による NDC 検索などが用いられている。しかし、事前にどのような件名の図書があるかを把握し、どのようなカテゴリー体系で管理されているかを把握していなければ、効率よく検索することができない。特に、専門外の分野や専門分野を学び始めたばかりの初学者の場合は、検索結果が数百件に上り、なかなか目的の資料に辿り着けないこともある。これに対し、図書館の司書との対話のように、対話形式での図書の検索支援により、漠然としたテーマから徐々に概念を絞り込み、図書検索に必要な検索クエリ—を推定することで、従来の書誌検索よりも効率の良い図書検索が行えると考えられる。

そこで、図書館内の蔵書に関する情報を書誌情報と目次からデータ化し、蔵書に関する知識を分散表現モデルとして構築する。また、これを対話形式での図書探索に用いることで、概念的に類似する専門用語による図書探索や図書の概念的な類似関係を考慮した新たな概念検索手法を確立することを本研究の学術的な問いとした。

２．研究の目的

本研究の目的は、図書の概念的な類似性をモデル化することで、学校図書館での司書との対話のようなレファレンスサービスを実現する対話型の図書探索手法を確立することである。学校図書館での図書探索では、利用者は図書探索に適した明確な検索ワードを熟知している状況は少ない。専門分野を学び始めたばかりの初学者の場合、漠然としたキーワードによって図書探索を行うために、目的の図書の探索ができなかったり、絞り込みに時間が掛かるという問題がある。また、概念的に類似する図書であっても、図書分類上は異なる分野として分類されていることがあり、分野を超えた関連図書の探索を難しくしている。

そこで、本研究では、システムとの対話形式での図書探索の過程で、対話文の中から専門用語を抽出し、利用者の図書の探索目的を類推し、対話の中で候補を提示しながら分野および図書の絞り込みを可能とする対話型の図書探索の実現手法を検討する。また、これを実現するために、書誌情報に含まれるタイトルと目次に出現する語彙の出現パターンから、専門用語および図書が扱うテーマを分散表現モデルとして構築する。そして、これを知識構造として活用することで、専門用語の類似概念の探索や学術図書の分野間の関係性を対話の中でフィードバックしながら、図書の綴り探索を実現するための探索手法を示すことを目的としている。

３．研究の方法

以下の３つの観点から研究に取り組むことで、直接的なキーワード検索だけでは得られない図書の探索を可能にするための概念検索手法を検討する。

（１）特徴語の意味空間の分析

利用者のクエリーに含まれる重要語を特定し、その類義を用いて検索クエリーを拡張するための特徴語の抽出手法およびその有用性を検証する。ここでは、既存の日本語 Wikipedia などをコーパスとして構築した単語分散表現モデルと比較して、図書情報を用いて単語分散表現モデルを構築した場合の類義語抽出と、それによる概念検索の有用性を検証する。

（２）専門書の分散表現モデルの構築

図書の内容としての類似性を考慮し、かつ、分野間の関係性を表現できるようにするための図書の分散表現モデルを構築し、その有用性を検証する。ここでは、図書のタイトルと目次の構成の関係に着目し、大規模言語モデルである事前学習済み Sentence-BERT モデルをファインチューニングすることで、目次の構成を文脈と見立てた分散表現モデルを構築する。これにより、図書の内容の類似性を分散表現モデルとして表現し、探索に活用できるようにする。

（３）概念検索手法の検討

対話ベースでの図書探索支援を実現するための概念検索手法を検討し、プロトタイプを構築してその有用性を検証する。図書館司書との対話のようにシステムとの対話の中で資料の綴り込みを実現するには、利用者から得た漠然としたキーワードから関連するキーワードや分野を類推し、利用者にフィードバックする必要がある。そこで、提案手法による検索クエリーの拡張と、絞り込みにより、従来のキーワードを中心とした検索手法では得られない、概念的に類似する図書をどれだけ探索できるようになるかを検証する。

４．研究成果

（１）特徴語の意味空間の分析

曖昧な検索クエリーを拡張するための関連語の抽出に取り組んだ。学術用語の専門書における概念空間モデルを構築するために、図書の目次データを用いて Doc2Vec による目次の文章ベクトルおよび単語の分散表現モデルを構築した。そして、既存の全文検索エンジンを利用して検索クエリーを探索し、ヒットした図書の目次に含まれる特徴語を抽出し、単語の分散表現モデルを用いて検索クエリーとの類似度を求めることで、利用者の検索クエリーに対して重要な関連語を抽出できるかを考察した。

その結果、目次情報をコーパスとして単語分散表現モデルを構築することで、既存の日本語 Wikipedia の情報をコーパスとして構築した単語分散表現モデルで得られる類義語よりも、検索クエリーが示す分野に特化した専門語が得られる傾向があり、絞り込み用の検索クエリー用のキーワード抽出に有用であると考えている。一方で、Doc2Vec による類似図書探索は、十分な精度が得られないことが分かった。

（２）図書の分散表現モデルの構築

概念的に内容構成が類似する図書の探索を実現するために、図書の分散表現モデルを構築した。ここでは、図書の目次構造をモデル化するために、文脈の関係をより捉えることのできる大規模言語モデルである事前学習済み Sentence-BERT モデルを用いて、目次の構造を文脈と見立てたファインチューニングをしている。Sentence-BERT で構築した図書の分散表現モデルを用いて図書の類似性を計算することで、分野横断型の図書探索への活用が期待できることを確認している。例えば、「画像認識」に関する図書を探索する場合に、医療情報を対象とした画像認識や、情報工学を対象とした画像認識の図書などが存在する。従来の図書分類ではこれらは異なる分野として離れた書架に配置されている。これに対し、構築した分散表現モデルによる類似度計算を行うことで、図書の内容としての類似度を考慮することが可能となり、分野間の横断的な関係が表現されていると考えられる。

（３）推薦アルゴリズムの検討

提案する対話型の図書推薦システムの概要を図１に示す。これは、対話形式によるキーワードの拡張と類似図書探索により、図書の絞り込みを支援する図書探索手法である。ユーザがシステムに対して、図書を探索する目的を入力すると、最初のステップとしてクエリー変換を行う。ここでは、対話文から抽出したクエリー文を基に、全文検索エンジンを活用して、クエリー文内に含まれる特徴語が含まれる図書を探索する。その後、全文検索でヒットした図書の分散表現ベクトルを基に、図書の内容として類似性に着目したクラスターを構築し、各クラスターの特徴語を抽出し、これらをユーザにフィードバックする。次のステップでは、ユーザが選択したクラスターを基に、クラスター内に存在する図書との類似図書を探索し、さらに、テーマごとにこれらを分類したものを推薦結果としてフィードバックする。

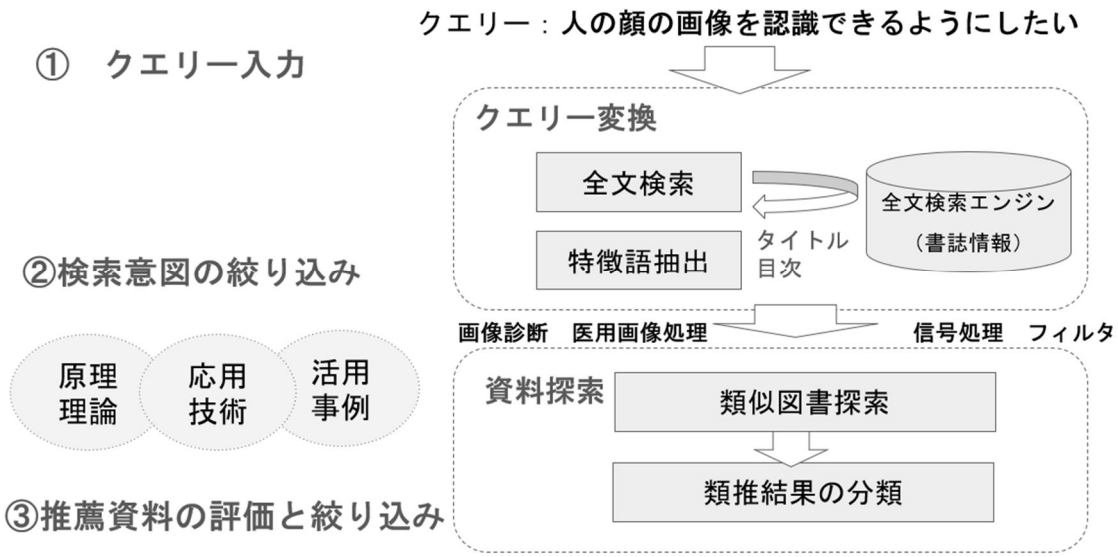

図１． 対話型の図書推薦システムの概要図

提案手法により、曖昧なクエリーからのキーワード抽出と、クエリーに関連するキーワードの分野および関連語の把握が可能になり、従来の全文検索では得られにくい関連分野を含む資料探索が可能になることを確認している。例えば、「西洋の建築の歴史について」知りたいという利用者のクエリーに対しては、図 2 の (a) に示すような「日本建築などとの構造面での違い」に重きを置いた図書や、「宗教や彫刻、絵画などの歴史や文化的な側面」に重きを置いた図書などが探索される。これらに対するユーザの反応を基に、再度、内容が類似する図書を探索することで、図 2 の (b-1) や (b-2) などのような目的に沿った分野の絞り込みを可能にしている。また、探索した図書の関係を分散表現モデルを用いて可視化することで、タイトルだけでは分かりにくい各図書の概念的な類似関係を把握できるようにしている。今後の課題として、学習者の資料探索の支援の実現に向けては、探索結果の提示方法や GUI の改善などの課題が残っているが、提案手法による図書の概念関係を活用した図書探索は、図書館司書の業務支援などへの活用にも期待できると考えている。

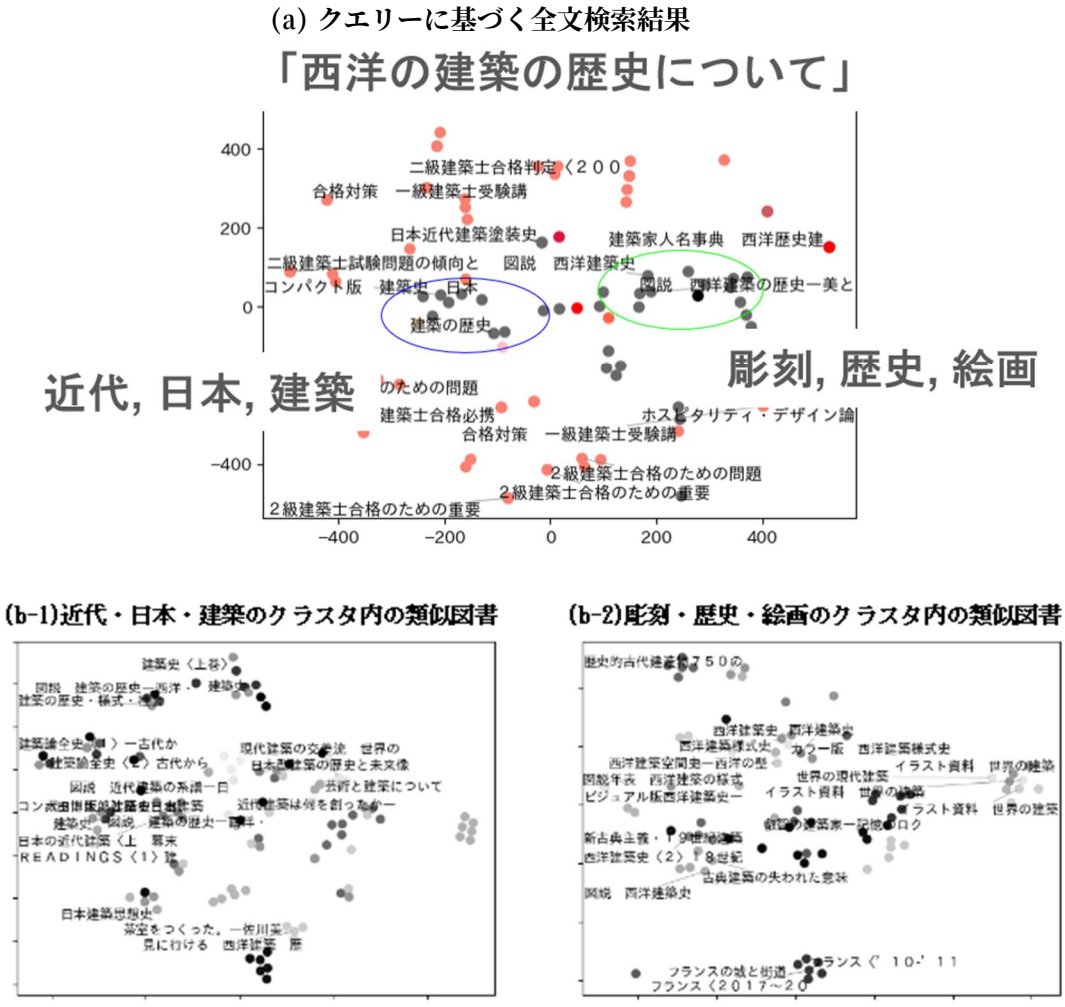

図 2.　検索意図に応じたクラスタ別の絞り込み例

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 関亜紀子
２．発表標題 Sentence-BERT を用いた対話型図書推薦システムの検討
３．学会等名 電子情報通信学会総合大会
４．発表年 2024年

１．発表者名 張冬旭，関亜紀子
２．発表標題 レファレンスシステムのためのキーワード抽出手法の一検討
３．学会等名 電子情報通信学会ソサエティ大会
４．発表年 2023年

１．発表者名 関亜紀子
２．発表標題 目次の分散表現による類似図書推定手法の一検討
３．学会等名 電子情報通信学会総合大会
４．発表年 2022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

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