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　５ 年　 ６ 月 ２０ 日現在

機関番号：１２３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１９８６

研究課題名（和文）作られる物と使われる物が記述された大規模文書群の検索に関する研究

研究課題名（英文）A Study on Information Retrieval for Large Document Corpora about Outcomes and Materials

研究代表者
安川　美智子（Yasukawa, Michiko）

群馬大学・情報学部・助教

研究者番号：７０３６１３８４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究は、情報検索の研究分野において重要な技術的課題の一つである「意味的に類似する索引語の綴りの不一致」に焦点を当て、特に、構造化された大規模テキストデータに内在する「成果物（作られる物）」と「材料（使われる物）」の意味的関係を抽出するための検索手法の開発を目的とする。研究期間内に、Institutional Research (IR)の研究において近年関心が高まっている研究情報や教育情報の戦略的な検索に応用可能な特徴分析の手法を開発した。また、IRの研究分野を対象に開発した特徴分析の手法を料理情報処理に応用し、日本の家庭料理の食材名や調理法を効率よく検索する手法を開発した。

研究成果の学術的意義や社会的意義
本研究は、日本語や英語のテキストデータにおける文字の並びや単語の並びを利用する従来法では対処できなかった「成果物と材料の意味的関係」のデータ構造を効果的に捉えるための新たな検索技術を開発したという点に学術的意義がある。また、大学における研究と教育、および、家庭における調理と栄養管理は、国内外において普遍的な重要性を持つ主題であり、これらを検索対象とする情報検索に応用可能な特徴分析の手法を提案したという点において本研究の社会的意義は大きい。

研究成果の概要（英文）：This study focuses on the spelling mismatch of semantically similar index terms. This spelling mismatch is an important problem in the research field of information retrieval. Therefore, this study aimed to develop a new technology for extracting the semantic relationship between "outcomes (what is made)" and "materials (what is used)", which often inherently exists in large structured text data. In this study, we developed a method of feature analysis that can be applied to strategic database searches of research and educational information, which has been of growing interest in Institutional Research (IR) in recent years. We also applied the feature analysis method developed for the research field of IR to culinary information processing, and developed a method for efficiently retrieving the lists of ingredients and their related cooking methods of Japanese traditional cooking.

研究分野：情報学

キーワード：情報検索　情報組織化　情報資源の構築・管理　自然言語処理　データベース

１．研究開始当初の背景

情報検索の研究分野における検索可能性の概念については、Onaifo[1]や Ivanovic[2]が、Search Engine Optimization (SEO)の観点から書誌情報の検索可能性を論じている。また、Larsson[3]は、博士論文（学位論文）の検索可能性について論じている。大学における教育情報の観点からは、シラバス分析については、Everly ら[4]が、一般的な授業用シラバスの属性情報や特徴をよりよく理解するために、シラバスを調査した結果を報告している。一方、授業シラバスのテキストデータ解析に関する先行研究の中には、人間の評価者を採用して分析している先行研究も報告されている。例えば，授業シラバスの文言の観点からシラバス評価を行った Ishiyama らの研究[5]，内容重視のシラバスと学習重視のシラバスの比較を行った Palmer らの研究[6]，講義における学生の必要性の観点からシラバスの主観的分析を行った Keller らの研究[7]がある．

２．研究の目的

本研究では、科学技術論文や授業シラバスのような、Relational Database(RDB)で管理される構造化された大規模テキストデータの検索可能性を議論する。具体的には、研究に関する情報（論文の書誌情報や研究費申請書）、教育に関する情報（大学図書館に所蔵されている教科書や参考書の書誌情報、学部・大学院の授業シラバス）、料理名や食材名を含む調理と献立に関する情報（テレビ番組や料理雑誌の公式サイトに掲載された料理レシピ情報）を対象として、文書解析的なアプローチによって、データ科学の観点から客観的にテキストデータ分析を行う手法を開発することを目的とする。

本研究では、以下のようなシナリオと研究的疑問（research question）を設定して、情報検索技術の開発と評価を行った。

(1) 数千以上の規模の文書が存在する場合、検索者は手作業で関連する文書を見つけることが困難となるため、コンピュータによる支援を必要とする。このような大規模な文書データベースにアクセスするためには、Text Mining (TM)などの計算機の支援を活用した文書処理が有用である。文書処理に機械学習手法を適用することで、一般的には大規模なテキストからの知識発見が可能となる。大規模なテキストを管理するためのデータベースは一種のテキストデータ（コーパス）であるため、知識を発見するには、従来の技術をもってすれば、簡単なことなのだろうか。言い換えれば、文書に従来のテキストデータ分析の手法を適用するだけで、知識獲得は自動的に成功すると考えて良いのだろうか。

参考文献

[1] D. Onaifo and D. Rasmussen, “Increasing libraries’ content findability on the web with search engine optimization,” Library Hi Tech, vol. 31, no. 1, pp. 87–108, 2013.
[2] L. Ivanovic, B. Dimic Surla, D. Surla, D. Ivanovic, Z. Konjovic, and G. Rudic, “Im- proving the discoverability of Ph.D. student work through a cris system,” The Elec- tronic Library, vol. 36, no. 3, pp. 471–486, 2018.
[3] J. Larsson, “The retrievability of a discipline: a domain analytic view of classifi- cation,” INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOUR- NAL, vol. 12, no. 4, 2007.
[4] M. B. Eberly, S. E. Newton, and R. A. Wiggins, “The syllabus as a tool for student- centered learning,” The Journal of General Education, pp. 56–74, 2001.
[5] J. T. Ishiyama and S. Hartlaub, “Does the wording of syllabi affect student course assessment in introductory political science classes?” PS: Political Science & Politics, vol. 35, no. 3, pp. 567–570, 2002.
[6] M. S. Palmer, L. B. Wheeler, and I. Aneece, “Does the document matter? the evolv- ing role of syllabi in higher education,” Change: The Magazine of Higher Learning, vol. 48, no. 4, pp. 36–47, 2016.
[7] C. E. Keller Jr, J. G. Marcis, and A. B. Deck, “A national survey on the perceived importance of syllabi components: Differences and agreements between students and instructors in the principles of accounting course.” Academy of Educational Leader- ship Journal, vol. 18, no. 3, 2014.

(2) 研究費の申請書や授業シラバスなど、文書が構造化された様式で記入されて、記入欄がいくつかある場合を考えてみる。記入欄は、「特に無し」等の意味のない文言ではなく、当該記入欄にふさわしい内容が記述されていることが望ましい。また、年度ごとに更新される文書の最新版には、以前のバージョンよりも検索可能な情報が含まれており、最新版の方が検索者にとって見つけやすい情報が含まれている方が、文書がより有効に検索されて活用されることが期待できる。では、文書の検索性を高めるには、どのような単語やフレーズを入れるのが効果的なのだろうか。また、文書に記載する単語をただ増やせば、検索可能性を高めることができるのだろうか。

(3) ある状況では、文書の執筆者が理想的な文書を自発的に書くことはできないが、システムによって関連する単語が提案されるならば、文書に追加する単語を選ぶことができると仮定する。この場合に、執筆者が多忙で、いくつかの単語提案しか読めない場合、単語提案を得るための外部のテキストリソースとして、どのようなデータベースを利用することが有効なのだろうか。

(4) 別の状況では、文書の執筆者はそれほど忙しくなく、多くの関連語を読むことに熱心である。この執筆者が有用な単語リストを得るためには、どのような提案が効果的だろうか。言い換えれば、検索されて活用される理想的な文書を執筆するために、何をどのように提案すれば検索可能性を高めることができるのか。

３．研究の方法

(1) 基盤的技術の開発

教育情報を対象として、大学の授業シラバスのデータを用いた手法の検討と評価を行った。具体的には、文書の検索可能性を調べるために、図１に示すような単語と文書を用いた情報検索の実験を行った。準備として、まず、シラバス文書コーパスに含まれるすべての文書に対して形態素解析を行い、全文検索のための転置インデックスに登録した。そして、シラバス文書コーパスから文書を１つ選び､その文書からクエリサンプルを得て､得られたクエリを用いた検索を行い、上位 n 位の文書リストを得た。

次にシラバス中の特定の単語の有効性を調べるために、「対象文書が k 位であった」ことを意味する Rank@k によって測定される検索有効性を学部ごとに調べ、タイトルによる検索と，各シラバスで最も特徴量の多い単語を用いたクエリ拡張を示す QE1 との相関を行列で表現する。行列において、各列と行は、それぞれタイトルによる検索と QE1 による検索の成功率を示している。各要素は、関連する列と行の対応関係である相関を表している。より具体的には、各要素の正規化値 nvxy を以下の計算式で得る。

$$nv_{xy} = \frac{\log(1 + v_{xy})}{\log(1 + v_{max})}.$$

例えば、行列の検索回数が 1,000 回の場合、行列要素の検索回数が 900，90，9，1，0 回である場合を考えてみる。上式により、これらの値はそれぞれ 1.000，0.663，0.338，0.102，0 として正規化される。行列の対角線上の要素は、タイトルと QE1 による対応する検索が同じ検索成功率であることを示す。例えば、左上の要素は、タイトルによる検索と QE1 による検索がともに成功し、対象の文書が検索結果の１位になったことを示し、対角線の要素の値の大きさは、QE1 による検索が改善されたとを示す。k が大きくなりすぎると、ユーザーは、検索結果を見なくなってしまう。そのため、対象文書が上位に表示されない検索を ″Failure ″とする。行列の可視化では、対象文書が k（k 10）でランク付けされた検索を ″ Success ″とし、それ以外を ″Failure ″とする。タイトル検索で失敗した結果の一部は″Success ″に移行する傾向にあるが、教養科目の授業シラバスの検索では、行列の対角要素の値が大きくなり、同じ文書の順位が維持されているケースが多い。教養の講義科目は１年生以上の受講科目であり、英語、数学基礎など同一または類似の科目が異なる講義科目で提供されている。そのため、クエリ拡張は、他のものに対するクエリ拡張に比べ、あまり効果がなかったと推察される。

逆に、学部専門科目の授業では行列の上部と右側の要素で検索成功率が大きな値となった。文系の大学院の授業タイトルの検索は、クエリが曖昧で対象文書のランキングを向上させるこ

とができず不成功に終わったが、（タイトルに限定しない）シラバス文書全体から得られた単語は有効な文書識別子であり、検索結果を向上させることが確認できた。

(2) 応用的技術の開発

① Google の Web 検索 API を利用して、料理に用いる食材や調理法に関連する検索キーワードから得られた URL を使用し、10 万件以上の日本語の Web ページを収集した。これらのデータは、将来のデータ分析や検索評価実験で活用できる。さらに、料理情報処理に関する国内外の研究状況や情報学分野全般の文献調査のために、関連語推薦の検討を行った。料理や食に関連する英語のキーワードを含む論文を DBLP で検索すると、過去 10 年間で論文数の増加が顕著であることが確認できた。近年、料理や食に関する情報処理の研究が重要性を増し、さまざまな研究領域の研究者が分野横断的なテーマに取り組んでいる。そこで本研究では、料理情報検索の手法を科学技術情報の検索など、他のドメイン（分野）の検索にも応用する検索応用を検討した。例えば、料理レシピの文書コレクションと論文書誌情報のデータベースには類似点があり、両方の検討を行うことで開発する技術の汎用性を高める効果が期待できる。また、情報学分野の論文書誌情報 DBLP と研究助成金申請書データベース KAKEN の横断検索において、情報検索フレームワークの検討を行った。

② 日本語の料理と食に関する Web ページの分析を行い、献立作成支援における情報検索手法を検討した。また、大学の教育 IR の観点から教育関連情報の分析にも取り組んだ。具体的には、料理レシピの大規模コーパスと食品名と栄養成分の対応関係を定義したデータベースを使用し、全文検索と集合演算を組み合わせたデータ分析を行った。また、大学図書館の蔵書情報と研究力の高い大学の授業シラバスデータを利用し、教育改善に有用な情報をテキストデータの分析により獲得する手法を提案した。これらの研究成果を国際会議で発表し、Outstanding Paper Award を受賞した。

③ クラウドソーシングを利用した単語概念の調査と大学の図書館所蔵情報を用いた時系列的な推移分析の手法を検討した。具体的には、クラウドソーシングによる調査で得られた知見をもとに、料理名、料理の写真、食材名の関係を調査した。収集されたデータを将来の研究でも活用する予定である。また、大学の学問分野の変化を把握するために、大学の所蔵書籍に基づいて 5 つの機械学習のアルゴリズムを用いた分析を行い、その結果を国際会議で報告した。

④ 論文著者や書籍の著者という観点から、特に他の研究者を超越する業績を持つ研究者を調査するための情報検索手法について検討した。このような検討のためのオープンデータが利用可能である一方で、効果的な情報検索手法は未だに不十分であるという問題がある。そこで、本研究では従来の単純な情報検索手法では対象データの特性上、効果的な検索が難しいケースを対象に、注目に値する教員・研究者を他の研究者と区別するためのデータベース検索手法を提案した。この提案手法は、複数データベースの連想的な検索にスカイライン演算を適用するものであり、外れ値が含まれる数値属性や変数間の負の相関がある場合に有効である。具体的には、科研費データベース KAKEN と大学図書館の蔵書データベース CiNii Books の 2 つのデータベースを使用し、教員・研究者の執筆数や予算額といった数値的な属性を評価した。また、得られた成果を国際会議 DSIR2021 で発表した。さらに、発展的な研究の準備として、異なる性質を持つ複数のデータベースを横断的に検索する方法についても検討した。

⑤ 科学技術文書の情報検索などの実用化が求められる分野への提案手法の応用について検討した。具体的には、大規模な研究助成金申請書データベースから収集した研究課題名を分析対象のテキストデータとして、データセットの分割と回帰分析を組み合わせた新しい手法を提案した。この提案手法は、教師あり学習によるパターン認識であり、インスタンスの類似度を用いている。従来の計算機環境では、計算量の問題から大規模なデータセットへの適用が困難であったが、最近のハードウェア、ソフトウェア、および仮想化技術の進歩により、実用的な時間でのデータ分析が可能となっている。評価実験により、提案手法がベースラインと比較して高い精度であることが確認できた。また、後半では、Institutional Research (IR)の研究手法を料理情報処理に応用し、日本の家庭料理のレシピデータを用いた評価実験を行った。文書類似検索と文書自動分類を組み合わせた特徴分析により、高精度な多クラス分類が可能な小規模なレシピデータセットを自動構築し、また、得られた類似文書の特徴量を用いて紙媒体の索引に相当する単語の一覧を作成することができる。これにより、情報アクセスの効率化と検索有効性の向上が期待できる。

４．研究成果

研究では、技術的な課題として、文書の検索可能な情報を測定し、文書に含めるべき効果的な単語を提案する方法を示す概念実証を行った。評価実験では実際のシラバス文書を用いて綿密な分析を行い、シラバス文書の検索可能情報に関する新しい知見を発見した。具体的には、目的で述べた研究的疑問を解明して、以下の答えを得ることができた。

(1) 大規模なテキストデータベース（文書コレクション）からの知識発見は、非自明なタスクである。従来のテキストマイニングのためにタスクに確立された手法を単純に適用するだけでは、問題を自動的に解決することはできない。

(2) 検索に有効な特徴語が一つであっても、検索可能な情報の一部となることができ、検索効果を高めることができる。授業シラバス中の単語の質を考慮せずに闇雲に文書の量を増やすことは、検索可能性を高める上で、現実的な解決策とはならない。

(3) 単語候補は、語彙の特性上、一般的な知識ベースよりも、大学図書館の書誌データベースなど、適合度の高いテキストデータベースから取得することが望ましい。

(4) より多くの単語候補を提案することが望ましい場合、単一の単語候補よりも、複合的な単語候補の方が、効果的な提案となる。

本研究における主な貢献は、研究目標を達成するために信頼性の高い方法を選択し、組み合わせる方法を実証したことである。本研究で得られた知見を、実際の教員の能力開発(Faculty Development)に応用することは今後の課題であるが、実験結果に基づき、より質の高い文書を執筆するための指針として、以下のことを提案する。

文書を検索できるようにすると、情報へのアクセス性が向上するが、研究費申請書やシラバスの登録フォームから詳細な情報を入力するのは手間がかかる。このため、内容の入力のための Web ユーザインタフェースを提供する場合、自動入力提案と組み合わせることが望ましい。このような入力補助の仕組みは、漢字の変換ミスなどのヒューマンエラーを減らすのに役立つ。また、Neologd などの形態素解析器のシステム辞書には最近の新語も含まれているが、言葉の表現や使い方によっては、誤って単語を分割してしまうことがある。シラバスの編集過程で、シラバス情報をデータベースに登録する前に、単語の選択について再考し、自動分類や情報検索の有効性を確認することは、文書の執筆者にとって有益である。

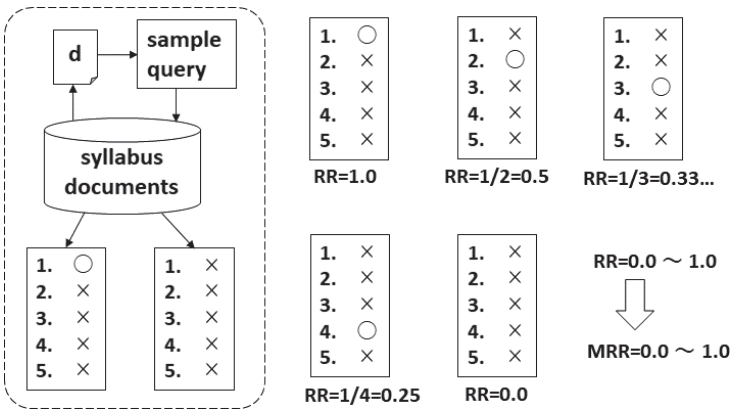

図１　検索可能性の評価

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2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 M. Yasukawa, H. Yokouchi, K. Yamazaki	４．巻 Vol.4 No.1
２．論文標題 Syllabus Mining for Analysis of Searchable Information	５．発行年 2020年
３．雑誌名 International Journal of Institutional Research and Management (IJIRM)	６．最初と最後の頁 46-65
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 安川 美智子	４．巻 34
２．論文標題 料理レシピ検索の評価用情報資源の構築	５．発行年 2019年
３．雑誌名 人工知能	６．最初と最後の頁 24-31
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10件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 *M. Yasukawa, K. Yamazaki
２．発表標題 Retrieval of Notable Academic People by an Ameliorated Skyline Operator
３．学会等名 Proc.of 10th International Congress on Advanced Applied Informatics (IIAI-AAI)（国際学会）
４．発表年 2021年

１．発表者名 *安川美智子
２．発表標題 研究年数と著者順序を考慮した論文生産性の可視化
３．学会等名 第10回 大学情報・機関調査研究集会（MJIR2021）
４．発表年 2021年

１．発表者名
安川美智子

２．発表標題
情報学分野を対象とした研究費獲得状況と図書館所蔵情報の横断検索システムの開発とそのFDへの応用

３．学会等名
第9回 大学情報・機関調査研究集会 (MJIR2020)

４．発表年
2020年

１．発表者名
M. Yasukawa, K. Yamazaki

２．発表標題
Categorizing Bibliographic Data for Detection of Transition in Academic Subjects

３．学会等名
Proc. of 8th International Congress on Advanced Applied Informatics (IIAI-AAI) (国際学会)

４．発表年
2020年

１．発表者名
*Michiko Yasukawa, Hirofumi Yokouchi, Koichi Yamazaki,

２．発表標題
"Syllabus Mining for Faculty Development in Science and Engineering Courses",

３．学会等名
Proc. of 8th International Congress on Advanced Applied Informatics (IIAI-AAI), pp.334--341, (国際学会)

４．発表年
2019年

１．発表者名
*安川美智子

２．発表標題
"情報学分野における教育能力開発を目的とした論文書誌情報の自動分類",

３．学会等名
MJIR2019 第8回 大学情報・機関調査研究集会 論文集 (Proceedings of the Eighth Meeting on Japanese Institutional Research),

４．発表年
2019年

１．発表者名
安川 美智子

２．発表標題
情報学分野における文献調査のための関連語推薦

３．学会等名
電子情報通信学会 人工知能と知識処理研究会

４．発表年
2018年

１．発表者名
安川 美智子

２．発表標題
科学技術論文の検索をテーマとする理工学系の演習授業の事例報告

３．学会等名
第25回 大学教育研究フォーラム 個人研究発表

４．発表年
2019年

１．発表者名
*M. Yasukawa

２．発表標題
Systematization of Japanese Culinary Information via Similarity and Heterogeneity among Documents

３．学会等名
IPSJ SIG Technical Reports, Vol. 2022-DBS-175, No. 43, pp.1--6

４．発表年
2022年

１．発表者名
*M. Yasukawa, K. Yamazaki

２．発表標題
Feature Selection by Thematic and Temporal Distinction in Research Grant Applications

３．学会等名
IIAI Letters on Institutional Research (Proc. DSIR2022), Vol.1, LIR019, pp.1--13（国際学会）

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