

哲学・倫理学文献のデータ化の有用性とテキストマイニングを用いた文献検索の見通し

土持, 貴志
九州大学大学院人文科学研究院

<https://doi.org/10.15017/7178796>

出版情報 : 哲学年報. 83, pp.79-91, 2024-03-18. Faculty of Humanities, Kyushu University
バージョン :
権利関係 :

哲学・倫理学文献のデータ化の有用性とテキストマイニング を用いた文献検索の見通し

土 持 貴 志

はじめに

本報告では、哲学・倫理学の書籍および論文をデータ化することの有用性およびデータ化の現状での問題について述べる。

現在、九州大学倫理学研究室では、研究室内の書籍の一部をデータ化する作業を進めている。データ化の目的は大きく2つある。第一に哲学・倫理学の書籍は質の高いものでも書店での取り扱いが中止^①になってしまふことが多く、学生が気軽に読んだり手に入れたりすることが困難になりがちな現状に鑑み、教員が選定した教育機関所蔵の書籍をデータ化することで、学生の信頼できる書籍へのアクセスを改善することである。第二に、特に学部生の卒業論文執筆の際に参考とすべき文献への効率的な検索を可能にすることである。哲学・倫理学では研究領域が多岐にわたる。現状ではそれらの研究内容や議論同士の関係性を検索するシステムはなく、学生自身が一から資料にあたるか、教員や所属院生等に個別に質問するかという、地道な方法しか存在しない。そこで、学生の所属する研究室等で所蔵している文献をデータ化し検索可能な状態にすることで、研究に資する環境を構築することを目指す。Google Scholar等を利用してインターネット上から文献を探すことも可能ではあるが、その文献がどれほどの重要性

を持つているかは学部生では判断がつかないことも多い。選書したデータベースを研究室で用意することで、論文執筆の際の資料探しの第一候補とすることが研究論文執筆の一助となると考えられる。

上記の目的を達成するために、次のフローでデータ化の作業を行っている。

- ① 倫理学研究室所有の資料の分類、整理を行う
- ② 資料のデータ化を行い、検索可能な資料にする
- ③ 資料に情報を付加する（キーワード、議論の構造）
- ④ 検索システムを構築する

本報告の構成は以下の通りである。第一節では、フロー①②の作業の現状を確認する。そして、Google Scholarと九州大学図書館といった他の文献検索サービスとの比較を行うことで、文献検索の点ではデータ化を行うことに利点があると述べる。第二節では、フロー③の資料に付加する情報とその活用として、キーワードの設定を挙げる。今後の見通しとして、テキストマイニングの手法を用いて各文献に機械的な処理を行うことでキーワード抽出を行うことの実現性と困難さについて述べる。第三節では、フロー③の資料に付加する情報のもう1つの例として、文献内に示される議論の構造をある程度機械的に取り出すことを挙げ、現状での可能性について述べる。

第一節 文献のデータ化の現状

現在のところ、和書約150件、洋書約400件の合計約550件の書籍のデータ化が完了した。基本的にはいずれの書籍も全文をデータ化し蓄積しており、これは詳細な文献検索を可能にするためである。他の文献検索サービスでは、文献そのものを所持していることは少なく、そのため文献を検索する場合にもタイトルや著者、要約で

の検索に限定されているが、全文検索を可能にすることで、より詳細な文献検索が可能となり検索性が向上する。論文執筆に役立つ哲学・倫理学の文献のみを集積できるシステムの構築が最終目標である。近年は応用倫理学のテーマで卒業論文を執筆する学生も多い。特に生命倫理のような学際的なテーマの文献を探す場合は、哲学・倫理学の文献にたどり着くことが困難になっている。本節では、中絶を例に Google Scholar、九州大学図書館、構築中のデータベースでどのような文献がヒットするかを比較する。

1. 文献の全文検索

文献をデータ化し本文を含む内容を所有することの具体的利点の1つは、特定の語が本文中に含まれるかどうかを複数のファイルを横断的に検索できる点である。他の文献検索サービスだと、その語がタイトルに含まれるか、目次やキーワードを別途設定しなければ検索できないことを考えると、全文検索が可能になることはデータ化の明確な利点であると考えられる。

全文検索は windows の標準機能で可能である。文献のフォルダにアクセスして、検索窓に検索キーワードを入力する。その後、検索窓下の「検索オプション」↓「ファイルコンテンツ」にチェックマークを入れることで文字認識を行った pdf や Microsoft Word 等の文章ファイルに検索ワードが含まれるかの検索が可能になる(図1)。コンピュータにローカル保存されたデータの検索に限られるとはいえ、有用な検索方法といえる。しかし、データ化した文献ファイルの本文に無制限なアクセスを許可することには著作権の問題がある。そのため、実際に検索システムを構築する際はファイルへのアクセス制限や、検索がヒットしたページのみを表示に制限するなど、著作権に抵触しない方式を採用する予定である。



図1 “abortion”がタイトルには含まれないが本文中に使用されている文献を検索できる

2. 他文献検索サービスとの比較

本節では、Google Scholar と九州大学図書館での検索結果と比較する。この2つのサービスを選んだ理由は、九州大学の学生が文献を検索する際に利用が予想されるものだからである。実際に論文執筆の資料を検索する場合を想定して“abortion ethics”で検索を行った。それぞれの検索結果は図2と図3の通りである。

Google Scholar の特徴は文献の網羅性にある。あらゆる分野の文献が収集されているため、膨大な文献が検索にヒットする。しかし、3つの問題点がある。1つ目は、哲学・倫理学以外の分野の文献も検索にヒットしてしまう点である。特に中絶のような学際的な主題では、他分野の文献のヒットが避けられない。図2の4件の文献の上でも上から2番目は医学分野の論文である。“abortion ethics”の検索結果の上位20件では、哲学・倫理学分野12件、法学分野4件、医学分野3件、社会学分野1件であった。問題点の2つ目は、文献の評価が困難である点である。網羅的に文献がヒットするので、その文献が哲学・倫理学の内容か等の評価は学部生には困難である。問題の3つ目は、検索結果から本文へアクセスする方法が多岐にわたり複雑化している点である。Google Scholar の検索結果は文献がオープンアクセス化しておりweb上で全文閲覧できるものや、文献の一部を公開しているものもある。上記の“abortion ethics”で検索した場合は4件が一部ないしは全文が公開されていた。現在では文献のオープンアクセス化が進

ち1番目はイスラム文化、3番目は日本思想に関する書籍である。“abortion ethics”の検索結果の上位20件を見ると、哲学・倫理学分野13件、法学分野3件、イスラム文化分野2件、歴史学分野2件であった。また、文献がE-journalでない限りは紙の書籍の確認が必須となるため、貸出中や別置の文献である場合は内容の確認に時間が必要になってしまう。さらに、九州大学図書館の検索サービスではシステム上に本文そのものを保持している訳ではない。そ

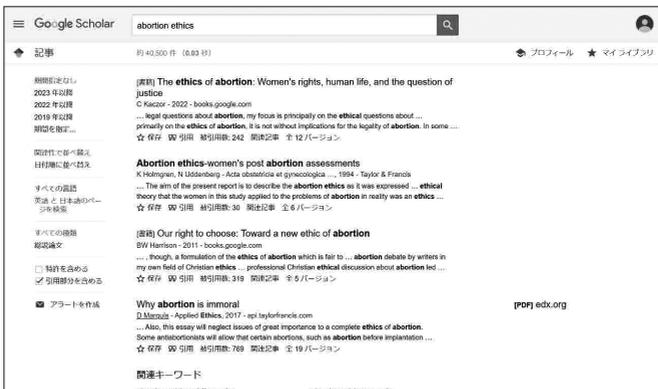


図2 Google Scholarでの“abortion ethics”の検索結果

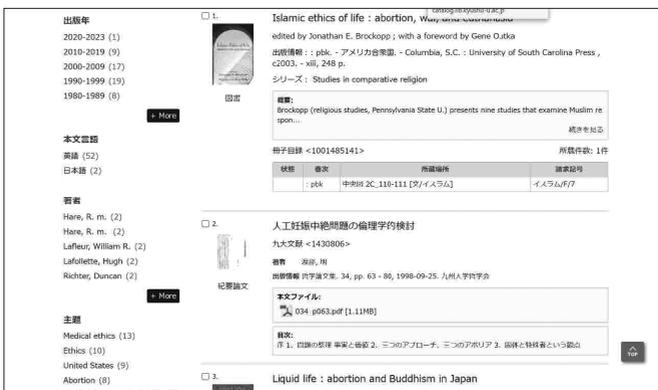


図3 九州大学図書館での“abortion ethics”の検索結果

んでいるとはいえ、多くの文献で検索可能なのは書誌情報のみであり、本文を確認するには別途文献を購入できるか調べ、取り寄せるといった手続きが必要であり即応性が低くなってしまう。九州大学図書館では選書という過程を経ている分、Google Scholarに比べて資料の信頼性は高いと考えられる。しかしながら、こちらはいくつかの問題点がある。九州大学図書館での検索でも他分野の文献のヒットは避けられない。図3の3件のう

のため、文献の検索はタイトルや著者名での検索に限定されており、全文検索を行うことはできない。

第二節 キーワード検索の実現性と困難さ

全文検索の利点は、複数の文献に渡って内容を網羅的に検索できることである。しかし、本文中に一度でもその語が使用されていれば検索結果にヒットしてしまうため、例えば中絶が主題でないものも、文献中に「中絶」が登場するならば検索結果に引っかかってしまうことになる。検索結果をよりコントロールするためには、それぞれの文献にキーワードを設定し、それによって検索することが有用である。本節では、文献のキーワードを機械的に抽出する方法の実現性とその難しさを述べる。その後、実際にキーワードをpdfファイルに設定する方法を示す。

1. KH Coder を利用したキーワードの抽出

文献のキーワードを機械的に抽出する方法としてテキストマイニングという手法を用いた。この手法を用いることで、文章中に特定の語が使用される回数や、特定の単語同士がセットで使用される回数を定量的に分析することができる。本報告では「KH Coder」を使用して、江口(2011)に掲載されている論文をいくつか分析した。この書籍を選択した理由は、中絶に対する賛成論、反対論、これまでの議論をまとめた論文がバランスよく所蔵されている点である。

抽出キーワードの例として中絶に反対する立場であるドン・マーキス「なぜ妊娠中絶は不道徳なのか」(図4)と、中絶を許容する立場であるマイケル・トゥーリー「妊娠中絶と新生児殺し」(図5)のキーワード抽出を行った。

図4と図5で抽出されているキーワードを比較すると分かるように、マーキス論文では「不正」や「殺す」とい

う語が上位にあり、一方でトゥーリー論文では「権利」や「原則」という語が上位にきている。使用されている語の頻度分析を行い、上位の語をキーワードとして設定することで、論文全体での話題に即した検索が可能になると考えられる。



図4 KH Coderによるドン・マークス「なぜ妊娠中絶は不道德なのか」のキーワード抽出



図5 KH Coderによるマイケル・トゥーリー「妊娠中絶と新生児殺し」のキーワード抽出

ただ、キーワード抽出を機械的に行うことは是非は検討されなければならない。キーワードを設定する際に最も信頼性が高い方法は教員あるいは院生等が実際に文献を読んで設定することであるが、時間的、経済的なコストを考えると現実的ではない。機械的にキーワードを抽出する方法は多量のファイルを短時間で処理できるため、時間的、経済的な問題は解決できる。その一方で、語の使用頻度は少ないが本文中の議論で重要な役割を果たしている語を見逃して、議論内容を捉えそこねてしまう危険性がある。

例えば、マークス論文には「避妊」という語が17回使用されており、キーワードの順位は54位である。しかし論文中では、胎児はわれわれと同じような将来の生を潜在的に持つから中絶は不正である、というマークスの主張に對して、同じく将来の生を潜在的に持つ配偶子を死なせる避妊も不正かという想定反論と、それに対するマークスの再反論が行われている。仮に避妊も不正であるという結論が得られるならば、自らの分析の難点になるとマークス自身が認めるように、避妊に関する議論はマークス論文の理解に重要な役割を持つ。単純に頻度でのキーワード抽出では分からない議論構造をいかに捉えるかが今後の課題である。



図 6 「文書のプロパティ」の開き方

2. キーワードを pdf ファイルに設定する方法

キーワードを pdf ファイルに設定するには、Adobe Acrobat Pro を用いた。キーワードを設定したいファイルを開き、「ファイル」↓「プロパティ」を選択し（図 6）、「文書のプロパティ」の「キーワード」の欄から設定が可能である（図 7）。

図 7 のキーワードを設定した白紙ファイルであっても、図 8 のように設定したキーワードでの検索が可能になる。

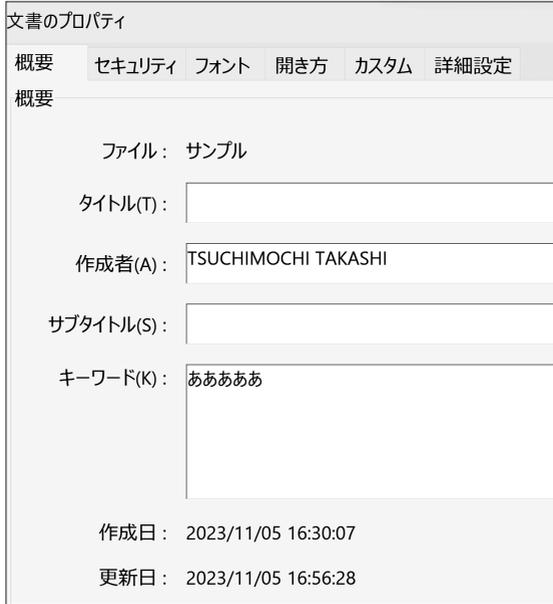


図 7 キーワードの設定方法

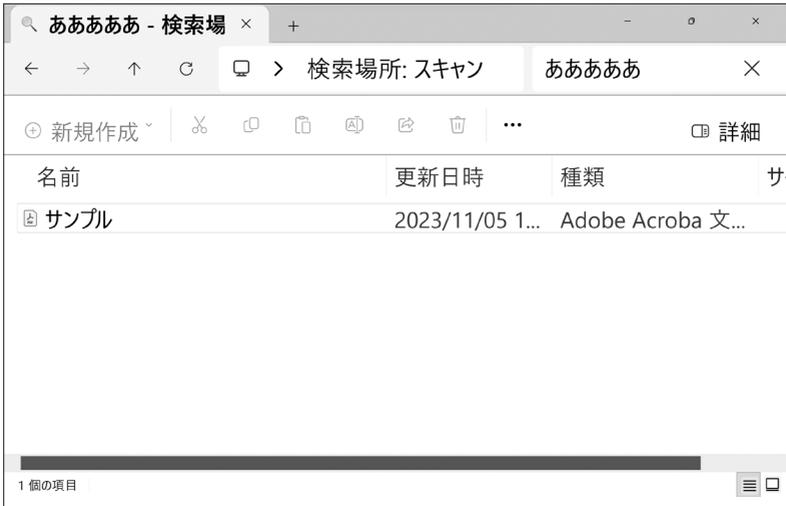


図 8 キーワード検索の結果

第三節 文献内の議論を機械的に把握することの困難さ

最後に、付加する情報として、あるいは検索可能な情報として望まれる、文献の哲学的・倫理的な議論の内容について、機械的に把握するという目標が現状ではどの程度達成できる見込みがあるかを述べる。文献のデータベースを構築する最終的な目標は、特定の文献が、どのような観点からのどのような議論によって、どのような立場を擁護あるいは批判しているのか、といった議論の構造が機械的に把握できるようになり、それを検索できるようになることである。

文献構造の機械的な把握の一例として、ChatGPTを用いた文章要約サービスを提供している「AIテキストマイニング」を用いて要約を行った。例えば、マークス「なぜ妊娠中絶は不道徳なのか」は次のように要約された。

本論は、中絶が道徳的に許容可能かどうかは胎児の道徳的地位に依拠するという前提の下で、中絶がきわめて不正であることを示す議論を提示している。胎児は、成人がそれをもっていればその人を殺すことが不正となるような性質をもっているのだから、中絶は不正である。本論は、宗教的な主張やローマ法王のドグマに依拠せず、論理的に健全であり、安楽死や避妊と矛盾しない。また、幼児に関する直観にも対応している。中絶の倫理にまつわる問題は、胎児の性質を決定するという問題であり、本論の議論はこの問題に解決をもたらすものである。^②

この要約では、マークスの主張とその根拠はうまく表されている。ただし、「宗教的な主張やローマ法皇のドグマ」に対するマークスの反論や、「安楽死や避妊と矛盾しない」とされる議論の構造は省かれてしまっている。

また、トゥーリー「妊娠中絶と新生児殺し」は次のように要約された。

この文章は、妊娠中絶と新生児殺しについての道徳的な問題を扱っています。著者は、人間の胎児と新生児が生命に対する権利を持つかどうかに関心を持っています。彼らは、有機体が生命に対する権利を持つためには、持続的な自己という概念を持ち、自分がそのような存在であると信じる必要があると主張しています。彼らはまた、新生児殺しの問題についても議論し、新生児が生命に対する権利を持たないと結論付けます。最後に、彼らは自己意識要件を擁護し、中絶と新生児殺しに関する保守的な見解を拒絶する必要があると主張しています。⁽³⁾

トゥーリー論文の要約も、マークス論文の要約と同じ特徴と問題を持っている。トゥーリーの主張とその根拠は表されている。ただし、「新生児が生命に対する権利を持たない」と結論付ける議論や、「中絶と新生児殺しに関する保守的な見解を拒絶する必要がある」と論じられる根拠は省かれてしまっている。

この2つの要約から見て取れる特徴に、①何の問題を扱っており、②何を主張しているか（どう結論しているか）をまとめている、という基本構造がある。この形の要約では、それぞれの論文の主題と筆者の主張を捉えることは可能だと思われる。しかしながら、この構造では、その主張がどのような議論構造によって導かれたのかを読み取ることは困難な部分も多い。これにより、哲学的な専門書の要約としては、有用性が限定的なものになっていると言えよう。というのも、哲学や倫理学の研究の多くは、「〜であるか」といった形の問いを提示し、これに対する答えを導くという形をとる。ここにおいて重要なのは、答えそのものではなく、答えに行きつくプロセス、すなわち論証の方である。なぜなら、1つには、答えはたいいていの場合、YesかNoかの大きく分けて二択になるので、答え

そのものが新しいとか示唆的であるということは稀なのである。また、経験的実証を行う他の研究分野と異なり、哲学の本質は答えでなく議論の構造の方にある以上、答えの部分自体は成果にはなりにくいのである。したがって、哲学的な文献を研究者や学生が検索する際には、もちろん「なんの問題について」（書籍の主題）「どう結論しているか」（筆者の主張）という情報も全く有用でないわけではないが、それらを「どのような観点から」もしくは「どのような議論によって」導いたのかの情報が必要ならば、あまり役には立たないのが現実である。

さらに、複数の反論に対する再反論を行うような論点を複数持つ論文はその議論構造を要約の形で取り出すことは一層困難である。例えば、ハーストハウス「徳理論と妊娠中絶」は次のように要約された。

この文章は、徳理論と妊娠中絶についての議論をまとめたものです。徳理論には、徳倫理学や徳に基づく倫理学といったさまざまな呼び方がありますが、これは義務論や功利主義理論に対抗できる理論として認識されています。しかし、徳理論にはいくつかの批判があります。この文章では、それらの批判について論じ、徳理論が妊娠中絶にどのように適用されるのかを考察しています。また、中絶に関する議論には女性の権利や生命の価値についての主張も含まれています。最後に、徳理論に対する批判や議論の結論についてのコメントが述べられています。^①

この要約からは、筆者が本文中で徳倫理に対する複数の反論を紹介し、それに対して再反論を行っている、ということを読み取れる。しかしながら、物語を起承転結で要約するように話の流れのみを抜き出した要約になっているため、具体的な議論の内容は省かれてしまっている。文献の議論の構造を機械的に抽出することは現時点では困難であると言えよう。

おわりに

本報告では、哲学・倫理学文献のデータ化の有用性と活用の困難さについて述べた。文献をデータ化する有用性は第一に絶版になり入手が困難になっている重要文献へのアクセスを確保すること、第二に複数の文献を横断的に検索可能にすることである。

文献検索の方法として、全文検索、キーワード検索、議論の構造の抽出を挙げた。全文検索は文章ファイルであれば windows の標準機能で可能である。キーワード検索は、テキストマイニングの手法を用いることで実装可能ではあるが、機械的な頻度順でのキーワード抽出の是非を検討する必要があることを指摘した。議論の構造を機械的に抽出することに関しては、AI による文章の要約を用いたとしても現状では困難であることについて述べた。

注

- (1) 例えば、J. ホスパーズ『分析哲学入門』や、M. スミス『道徳の中心問題』
- (2) IV テキストマイニングを用いたドン・マーキス「なぜ妊娠中絶は不道徳なのか」の要約
- (3) AI テキストマイニングを用いたマイケル・トゥーリー「妊娠中絶と新生児殺し」の要約
- (4) IV テキストマイニングを用いたロザリンド・ハーストハウス「徳理論と妊娠中絶」の要約

参考文献

樋口耕一、中村康則、周景龍『動かして学ぶ！はじめてのテキストマイニング』、ナカニシヤ出版、2022年
江口聡編・監訳、『妊娠中絶の生命倫理 哲学者たちは何を議論したか』、勁草書房、2011年