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　 ９ 日現在

機関番号：１２１０２
研究種目：研究活動スタート支援
研究期間：2021～2022
課題番号：２１Ｋ２１３０３

研究課題名（和文）学術文献の参照に着目したWikipediaにおける情報提供の信頼性の評価

研究課題名（英文）Reliability of Information on Wikipedia based on Scholarly Bibliographic References

研究代表者
吉川　次郎（Kikkawa, Jiro）

筑波大学・図書館情報メディア系・特任助教

研究者番号：８０９０８４００

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：本研究では，研究代表者が開発した手法を用いて，2021年10月時点の英語版Wikipediaにおける31万件の記事上の147万件の学術文献の参照記述について，それぞれの参照記述が初めて登場した時点（初出時点）を特定したデータセットを構築した．同データセットの分析を通じて，Wikipedia記事自体の新規作成から最初の参照記述が追加されるまでの時差が近年になるにつれて短くなる傾向があることを明らかにした．この傾向は2007年から2008年頃から顕著になっており，その当時に掲げられた「コンテンツの量から質への転換」を目指す方針に対応するWikipediaコミュニティの動きとして解釈できる．

研究成果の学術的意義や社会的意義
本研究では，世界で最も閲覧の多いウェブサイトのひとつであり、生命や健康に関わる情報を得る際にも利用されるWikipediaにおける情報提供の信頼性の評価として，学術文献に裏付けられた正確な情報の提供状況に関する分析を行った．
本研究で得られた知見は，Wikipediaの記事を通じて社会の誰もが学術知識に触れ，より良い意思決定を行ううえで，学術文献の参照記述がどのように有効活用されうるのかを検討した点で意義が大きい．また，Wikipediaの記事は継続的に加筆・修正が行われるため，学術文献の参照記述を追跡するための手法を構築し，プログラム等を公開した点においても学術的・社会的な意義がある．

研究成果の概要（英文）：In this study, we built a dataset of the first appearances of the 1.47 million scholarly bibliographic references on the 310,000 Wikipedia articles of scholarly references on English Wikipedia as of October 2021 by using our proposed method.
Moreover, we conducted the time lag analysis on this dataset and observed that the time lag between creating Wikipedia articles and adding the first references to the corresponding articles decreased over time in more recent years, particularly since 2007-2008. This trend can be seen as a response to the policy changes of the Wikipedia community at that time.

研究分野：図書館情報学

キーワード：図書館情報学　計量書誌学　学術情報流通　オープンサイエンス　Wikipedia　学際性　永続的識別子

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ウェブを通じた学術文献の閲覧・入手が定着し，研究者や専門家などの従来の利用者のみに留まらず，多様な人々やコミュニティによる学術文献の利活用が生じている．たとえば，オンラインのフリー百科事典である「Wikipedia」では，不特定多数の協働による編纂を行ううえで，信頼できる情報源として学術文献を参照することが推奨されている．実際に，研究代表者による調査を通じて，英語版 Wikipedia では 2017 年 3 月時点で 93 万件もの文献が参照されていることが明らかになっている．

Wikipedia は世界で最も閲覧の多いウェブサイトのひとつである．また，生命や健康に関わる情報を得る際にも利用されることから，学術文献に裏付けられた正確な情報の提供は不可欠である．加えて，新たな学術成果が公表され，知識や学説自体が移り変わっていくことを踏まえると，多数の文献を単に参照するだけでなく，継続的な学術文献の参照記述 (以下，文献参照) の追加や更新を行う必要があると考えられる．

しかし，これまで，文献参照が各記事に初めて追加された時点 (以下、初出時点) の特定自体が難しいことから，文献参照の現況の分析や課題の検討はほとんど行われてこなかった．

２．研究の目的

上述した背景に基づき，本研究では「Wikipedia での学術文献に基づく信頼性の高い情報提供の定量的な評価」を目指す．このような評価を実現するために，(1) Wikipedia 上の文献参照の初出時点を特定したうえで，(2)文献の新しさに基づく知識・学説の反映状況，ならびに，(3) Wikipedia 記事の作成から文献参照の追加までの時差に関する分析を行う．これらの分析を通じて，Wikipedia における文献参照の追加に関する現況把握および課題発見に貢献するとともに，学術文献に基づく知識への容易なアクセスを促進することを研究の目的とする．

Wikipedia 上の学術文献の参照記述に関する既往研究は，学界での評価指標との関係 (Impact Factor が高い雑誌の論文が参照されやすいのか)，オープンアクセス文献の参照状況 (全文公開の文献が参照されやすいのか) の調査などに留まっている．加えて，そもそも，任意の Wikipedia 記事におけるそれぞれの文献参照に関する初出時点の特定自体が困難であるため，分析対象の取得自体が実現されておらず，文献参照の現況の分析や課題の検討はほとんど行われていない．

以上の状況から，研究代表者はこれまでに文献参照の初出時点を高い精度で特定するための手法を構築し，文献参照およびその追加を行う編集者の分析を行ってきた．本研究課題では同手法を用いて文献参照の初出時点を特定したうえで，それらの文献参照の分析を行うことにより，Wikipedia の学術知識の信頼性に関する定量的な評価を実現する．

３．研究の方法

本研究における分析対象に関する前提として，「学術文献の参照記述」およびその「初出時点」について，それぞれ以下のように定義する．

まず，学術文献の参照記述については，特定の学術文献を一意に識別可能な Wikipedia 本文中の記述と定義する．したがって，論文の引用文献リストのように，著者名，文献タイトル，雑誌名，出版年，巻・号，掲載ページなどがセットで記述されていることを必ずしも前提とするものではなく，URI や特定の識別子のみが示されている場合であっても文献を一意に識別可能であれば対象に含める．なお，ここでの学術文献は主に学術論文を対象とするが，ウェブ上で流通している信頼性が高い学術情報が英語で書かれているとは必ずしも限らないため，使用言語については特に限定せずに扱うこととする．

次に，学術文献の参照記述の「初出時点」は，それぞれの参照記述が任意の Wikipedia 記事に新規追加された時点とする．

これらの定義を踏まえたうえで，以下に示す 3 つの課題に取り組む．

(1) 英語版 Wikipedia における学術文献の参照記述の初出時点に関するデータセットの構築

これまでに研究代表者が考案した方法論を用いることで，英語版 Wikipedia における学術文献の参照記述の初出時点を特定する．研究代表者の提案手法は図 1 に示すように，学術文献の識別子および文献タイトルを用いるものである．具体的には，Wikipedia の過去の記事本文をすべて取得して古い順に展開し，識別子 (DOI や PubMed などの文献 ID) が含まれるか，文献タイトルの全体または一部 (先頭 5 単語) が含まれるか，記事本文中に文献タイトルと類似度の高い文字列が含まれるかを調査し，いずれかの条件を満たす最も古い編集を検出する．

この手法を英語版 Wikipedia に適用して検証を行った結果，精度は全体で 93.3%，22 分野中 20 分野で 90%以上であり，分野を問わず概ね高い精度で初出時点を特定可能であることが明ら

かになっている．本研究では，2021 年 10 月時点での英語版 Wikipedia に提案手法を適用することで学術文献の参照記述の初出時点データセットを構築する．

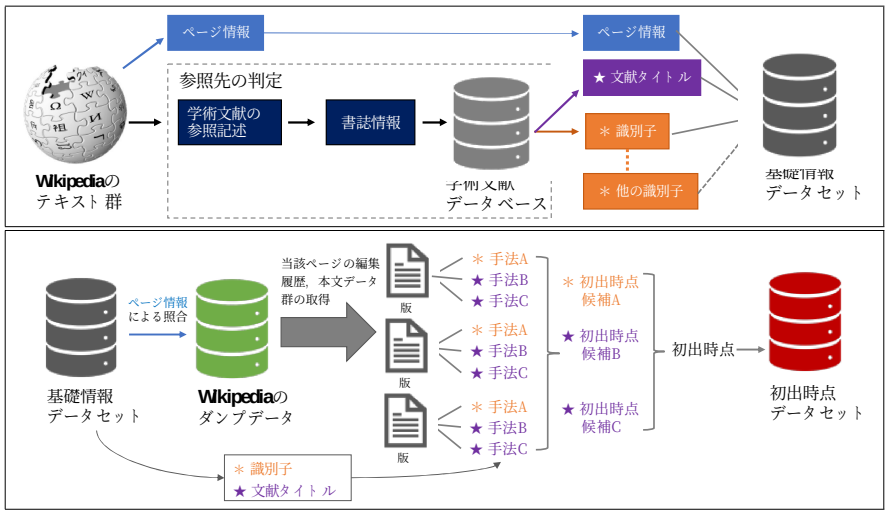

図 1: 提案手法の概要

(2) 文献の新しさに基づく知識・学説の反映状況の分析

前述した学術文献の参照記述の初出時点データセットを用いて，Wikipedia で参照されている学術文献の識別子から書誌情報を取得し，文献自体の公表年と Wikipedia 上での参照年における時差を算出する．その結果に基づき，作成年が新しい Wikipedia 記事において，より公表年が新しい学術文献が参照される傾向があるのかどうかを明らかにする．

(3) Wikipedia 記事の作成から文献参照の追加までの時差の分析

前述した学術文献の参照記述の初出時点データセットに含まれる Wikipedia 記事の作成日時の情報を取得・追加したうえで，それぞれの Wikipedia 記事の作成時点と，当該記事における文献参照の初出時点の時差を算出する．同一記事に複数の文献参照が存在する場合は，最初の 1 件が追加された時点を対象に算出を行う．その結果に基づき，Wikipedia 記事の作成年ごとにグループを分けたうえで，Wikipedia 記事の作成年が近年になるにつれて時差に変化が生じているのかどうかを明らかにする．

４．研究成果

(1) 英語版 Wikipedia における学術文献の参照記述の初出時点に関するデータセットの構築

研究代表者が開発した手法を用いて，2021 年 10 月時点の英語版 Wikipedia における 31 万件の記事から参照されている 147 万件の学術文献について，それぞれの初出時点を特定したデータセットを構築した．

さらに，同データセットに関するデータ論文を Scientific Data 誌にて公表することを通じて，当該データセット本体だけでなく，その構築に必要なツール群を公開した．このデータセットを分析対象とすることで 2021 年 10 月時点での英語版 Wikipedia における文献参照の分析が可能であることはもちろんのこと，これらのツール群を用いることにより，任意の時点での Wikipedia における学術文献の参照記述の初出時点を特定することができる．

(2) 文献の新しさに基づく知識・学説の反映状況の分析

2021 年 10 月時点の英語版 Wikipedia における学術文献の参照記述の初出時点データセットを用いて，文献自体の公表年と Wikipedia 上での参照年の時差を算出した結果，Wikipedia の記事が作成された年に関わらず，時差の中央値は 6.0 から 7.0，最頻値は 0 (単位は年) であった．このことから，作成年が新しい Wikipedia 記事において，より公表年が新しい学術文献が参照される傾向は見られない．

なお，最頻値に関しては 2007-2008 年のみ 5 であり，他の年と異なる傾向が見られたが，これは一定の編集作業を機械的かつ自動的に行うプログラムである Bot により，細胞生物学・分子生物学分野における 2 編の学術文献の参照記述が多数の記事に追加された影響によって生じた結果である．

以上の研究成果を国際会議「iConference 2023」にて発表した．

(3) Wikipedia 記事の作成から文献参照の追加までの時差の分析

2021 年 10 月時点の英語版 Wikipedia における学術文献の参照記述の初出時点データセットを用いて，それぞれの Wikipedia 記事の作成時点と，当該記事における文献参照の初出時点の

時差を算出した結果，近年になるにつれて時差が短くなる傾向があることが明らかになった．特に，Wikipedia の記事作成と同時に参照記述が追加されるケースは，2007 年から 2008 年にかけて急増後，2013 年以降は過半数を占めるようになっている．

これらの結果は，Wikipedia のコミュニティにおいて学術文献の参照記述を迅速に追加するという習慣が広まっていることを示すものであるといえる．また，2007 年から 2008 年にかけて見られる急増については，Wikipedia の共同創始者であるジミー・ウェールズ氏が 2006 年に「コンテンツの量から質への転換を目指す」という方針を掲げたことに対応する Wikipedia コミュニティの動きとして解釈することができる．

以上の研究成果を国際会議「iConference 2023」にて発表した．

上述した研究成果に加えて，本研究では以下に示す課題にも取り組んだ．

(4) Wikipedia 上での学術文献の参照記述の追加活動が分野横断的 (学際的) であることからアイデアを敷衍させ，学術ビッグデータを用いて研究活動 (研究成果としての論文発表) における学際性の動向分析を行い，複数の研究分野を対象に，それぞれの特徴や経年的な変化を明らかにした．この研究成果を国際会議「A-LIEP 2021」ならびに国際英文誌「LIBRES: Library and Information Science Research e-Journal」にて発表した．

(5) 学術文献の識別子 (デジタルオブジェクト識別子，DOI) 自体の正確性と安定性に関して，DOI の大規模データセットを用いた分析を行った．その結果，約 70 万件の削除済の DOI を特定した．さらに，これらの DOI について，コンテンツの種別や削除の発生要因を明らかにした．これらの分析結果について，国際会議「TPDL 2022」にて発表を行った．

本研究課題の遂行を通じて，Wikipedia の学術知識の信頼性に関する定量的な評価を行うための基礎的なデータを収集・分析し，その結果に基づく有用な知見を獲得することができた．ただし，より精緻な分析や評価を実現するためには，それぞれの文献参照に関する類型・種別を区別したり，研究分野や領域ごとの文献参照の特徴を明らかにしたりする必要があると考えられる．これらの点については，今後の研究課題として積極的に取り組んで行きたい．

５．主な発表論文等

〔雑誌論文〕　計3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　2件）

１．著者名 Kikkawa Jiro, Takaku Masao, Yoshikane Fuyuki	４．巻 9
２．論文標題 Dataset of first appearances of the scholarly bibliographic references on Wikipedia articles	５．発行年 2022年
３．雑誌名 Scientific Data	６．最初と最後の頁 1~11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41597-022-01190-z	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 Takei Chizuko, Kikkawa Jiro, Yoshikane Fuyuki	４．巻 32
２．論文標題 Progress in Interdisciplinarity: Bibliometric Analysis of the Diversity of Researchers’ Fields of Specialization Over a 20-Year Period	５．発行年 2022年
３．雑誌名 LIBRES: Library and Information Science Research e-Journal	６．最初と最後の頁 64-80
掲載論文のDOI（デジタルオブジェクト識別子） 10.32655/libres.2022.1.5	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 吉川 次郎	４．巻 38
２．論文標題 オープンな学術情報がもたらす新たな知識基盤: Wikipedia上の学術文献の参照記述に関する研究動向を中心に	５．発行年 2023年
３．雑誌名 人工知能	６．最初と最後の頁 399-407
掲載論文のDOI（デジタルオブジェクト識別子） 10.11517/jjsai.38.3_399	査読の有無 無
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3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Takei Chizuko, Kikkawa Jiro, Yoshikane Fuyuki
２．発表標題 Progress in interdisciplinarity: From the perspectives of diversity of researchers' fields of specialization
３．学会等名 The 10th Asia-Pacific Conference on Library & Information Education and Practice (A-LIEP 2021)（国際学会）
４．発表年 2021年

１．発表者名
Kikkawa Jiro, Takaku Masao, Yoshikane Fuyuki

２．発表標題
Analysis of the deletions of DOIs: What factors undermine their persistence and to what extent?

３．学会等名
Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022), Lecture Notes in Computer Science (LNCS)（国際学会）

４．発表年
2022年

１．発表者名
Kikkawa Jiro, Takaku Masao, Yoshikane Fuyuki

２．発表標題
Time Lag Analysis of Adding Scholarly References to English Wikipedia: How Rapidly Are They Added to and How Fresh Are They?

３．学会等名
Proceedings of the 18th International Conference, iConference 2023, Lecture Notes in Computer Science (LNCS)（国際学会）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
(1) Dataset of first appearances of the scholarly…
https://doi.org/10.5281/zenodo.5595573
(2) corgies/sdata2021: v1.0 | Zenodo
https://doi.org/10.5281/zenodo.5776204

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