研究指導　中澤 真 准教授

Wikipedia 記事の信頼性に関する一考察 —参照情報がもつ多様なデータからの信頼性評価—

唐橋　慧

1.　はじめに

近年，認知度の高いオンライン百科事典としてWikipediaがある．2016年12月現在，日本語版には100万件を超える記事が存在し[1]，今後も記事数は増えていくと思われる．一方でWikipediaはしばしばその記事の信頼性を問われている．それはWikipediaの方針のひとつに「真実かどうかではなく検証可能かどうか」という項目があるからである．つまり，査読されていない記事をユーザーが読み，自身で正誤を判断しなければならないということである．言葉の意味を調べているにもかかわらず，初めて知る言葉の意味を自分で判断することは難しい．

これに対して，Wikipedia記事の信頼度を示すことでユーザーの判断を助ける信頼度算出式が考えられてきたが，信頼性の評価指標としては不十分な点もまだある．そこで本研究では，リンク切れ[1]や誤記述情報[2]が信頼度に含まれてしまう問題を解決する．そのため，Wikipedia記事が含む参照情報を元に新たな信頼度評価指標を考察する．

2.　Wikipediaの現状と課題

2.1　Wikipediaの信頼性

Wikipediaは気になった言葉の概要を知るには便利なサイトであり，日本語版Wikipediaには100万件を超える記事が蓄積されている[1]．これほどの記事があるのは，匿名で誰でも投稿や編集が可能という特徴をWikipediaが持っていることに起因する．一方，ユーザー側から考えると，記事の内容が信頼できるものであることが望ましいが，情報のすべてを信用する人は多くない．

Wikipediaの利用動向を調査した長塚ら[2]は，鶴見大学の学生を中心にWikipediaの認知度や利用経験，目的，信頼度等の項目を含むアンケートを実施した．その結果，認知度と利用経験ともに96.6%，信頼度は高校生，学生，司書講習受講者の順に40.0%，26.6%，2.3%というアンケート結果を示している．このことから，Wikipediaの認知度は高いが，一方で掲載されている情報への信頼は得られていないことがわかる．

実際，Wikipedia記事は特定の機関等による事実関係の確認が行われないため，しばしば信頼性が問題視されている．いかなる者からも記事に対する評価が行われないために，その記事が正しいかどうかを判断するのは閲覧者自身となり，初めて知る言葉の正誤を判断することは難しくなる．このため，Wikipedia記事の信頼度の指標を何らかの形で閲覧者に提供することは，利用する者にとって正誤判断の有用な材料となる．

2.2　Wikipediaの信頼性に関する課題

Wikipediaの各記事の信頼性を判断するには，その記事の参照情報[3]は誰が記したものなのか，記事は間違いなく記載された参照情報を参考にして書かれているかの2つの点を確認しなければならない．前者の著者問題に関しては，Wikipediaだけでなくインターネット上の情報全体で共通の課題であり，2016年にはキュレーションメディア[4]における記事内容の信頼性問題ということで話題にもなった．この場合においても，その記事の執筆者がどのような立場の人間・組織なのかを明らかにすることが信頼性向上に必要だと指摘されている[3][4]．後者の参照情報問題については，Wikipedia記事内容と書籍やWeb情報等の参照情報に不一致な点があり，整合性が取れていない可能性があることが挙げられる．

参照情報が記載されていれば，ユーザー自身が参照先の執筆者や組織を確かめることにより，情報源としての信頼性を判断することもできる．また，Wikipedia記事に参照情報の内容が的確に反映されているのか確認することにより，記事そのものの正誤の判断も可能になる．ただし，参照情報がインターネット上の情報を指すURLであった場合には，リンク切れにより情報にアクセスできない可能性もある．このような場合には，記事内容の正誤に関する判断材料が失われてしまい，信頼性の評価も難しくなってしまう．逆に，参照情報のリンク切れは，その情報の正しさを担保する出典の喪失を意味し，記事内容の信頼性を下げる評価指標と考えることができる．後述する先行研究では，リンク切れを考慮せずに参照情報の数をカウントして信頼性に関する評価値を算出しているが，これでは正確な評価はできない可能性が高い．そこで本研究ではリンク切れを考慮しつつ，記事に含まれる参照情報が持

1　404 エラーが表示される場合のほかに参照情報に記載された URL にアクセスした際，ページが削除されている場合や URL が変更されてページを閲覧できない場合，本来の目的ページが表示されない場合をリンク切れと判断している．

2　ここでは，書籍内容を Wikipedia 記事に書き写す際に誤って入力した情報のことをさす．

3　ここでは，参照情報を Wikipedia 記事内に記された参考文献や脚注のこととする．

4　収集した情報を分類し，つなぎ合わせて新しい価値を持たせたコンテンツをもつメディアのことをさす．

つ多様な情報を用いた信頼性評価を行い，既存のWikipediaの信頼性評価手法の改善を図る．

3. Wikipediaの信頼性評価

Wikipedia記事の文末には，執筆者が参考にした書籍やURLが記載されている．書籍は記事内の参考文献に含まれ，URLは記事内の主に脚注に含まれる．本研究ではWikipediaの信頼度算出に，これらの参照情報をどのように用いるべきかを明らかにする．そこでまず，参照情報に基づいて信頼性の評価値を算出した既存の2つの研究について以下で述べる．

井上ら[5]はWikipedia記事に含まれる記載された参照情報として，参考文献と信頼できるドメインの数をカウントし，この値が大きい記事ほど信頼できるという仮定のもと，信頼度算出の評価式を示した．書籍や雑誌を含む参考文献は発行にあたり出版社等から複数回のチェックを受けているため，信頼度の高い情報とみなすことができる．一方で，Web上では誰が書いたものかがわかる情報からわからない情報まで玉石混交している．そこで井上らはWeb情報の信頼性について，URLに含まれるドメインによって判別する方法を提案した．信頼できるドメインには政府機関が使用する「go.jp」や高等教育機関の「ac.jp」，各新聞社が持つドメインがあると考え，これらを評価値の算出に織り込んでいる．このようにして算出した信頼度が，Wikipedia記事の信頼性を評価する値として一定の有効性があることを評価実験により示している．

井上らの研究の改善を図った吉次ら[6]は，Wikipedia記事に含まれる書籍のISBNの有無やURLのPageRank[7]スコアを用いることで，信頼度の精度向上を図った．井上らは書籍を一括してカウントしたが，吉次らはISBNがある書籍はISBNがないものより信頼度が高いと判断し信頼度に反映した．またWeb情報についてはすべてのドメインを対象にPageRankのスコアを信頼度算出の評価値として用いている．しかし，評価実験の結果では，井上らより良い結果は得られなかった．なお，PageRankは2016年4月から機能が廃止されているため，現在では利用することができない．

ここで述べたいずれの先行研究においても，信頼度を算出するために参考文献や脚注の参照情報を用いている点は共通している．しかし，これらの情報だけではリンク切れの問題や誤記述情報を信頼度スコアに含めてしまう問題があり，信頼度の算出方法にはまだ改善の余地がある．

4. Wikipedia記事内の参照情報の傾向分析

本節では，参照情報の傾向を把握するための調査方法とその結果について述べる．

4.1 調査対象データ

はじめに，調査対象とするカテゴリを決定する．Wikipedia記事はカテゴリ別に分類されており，9つの主要カテゴリがあり，その下に重要な100のカテゴリが連なる形で分類されている．本研究では調査対象として，新しい情報が次々と生まれてくることで更新頻度が高くなり，情報の電子化も進んでいるカテゴリとして「情報」，「医療」を選び，逆に紙媒体の資料が中心で電子情報が少なく，情報の更新頻度も少ないカテゴリとして「歴史」，「芸術」を選択した．4つのカテゴリに属する記事は用語辞典等からキーワードを無作為に抽出し，各カテゴリに属するキーワードのみを30記事分選び出した．次に選択したキーワードのWikipedia記事に含まれているすべての参照情報について，表1に示す9項目について調査した．

表 1:調査項目

No.	調査項目	概要
1	タイトル	参照情報のタイトル
2	資料発表年	その情報が公開された年
3	ドメイン名	記載URLのトップレベルドメインやセカンドレベルドメイン
4	著者名（組織名）	その情報を書いたのはどのような者か
5	リンクの有無	参照情報にリンクが設定されているか
6	リンク切れの有無	URLで指定された先に該当する参照情報が存在するか
7	記載ページ指定	ページ指定に誤りがないか
8	情報形態	a. 機械判読可能：オリジナルが電子データで作られている
		b. 機械判読可能：OCRによって電子化されたデータ
		c. 機械判読不可能：機械判読できない電子データ
		d. 機械判読不可能：書籍や雑誌
9	Wikipedia記事の整合性	不一致なし
		年号の不一致
		数値の不一致
		言葉の不一致
		内容の不一致

4.2 分析結果と考察

参照情報調査の結果，全120記事から798件の参照情報が確認され，そのうちWikipediaの記事内容と一致していない参照情報はわずか5件だった．表2は不一致箇所があった5件の参照情報を調査した結果である．5件とも電子データをオリジナルとした情報形態であり，不一致箇所の種類も数値データという共通の性質を持っている．

このことから，コピー＆ペーストが可能な電子データの情報形態でさえ，Wikipedia記事との間に不一致が生じるとわかった．一方で参照情報が書籍など機械判読不可能な場合には，書き写しの作業が必要となるため，電子データ以上の不一致が生じると考えられる．先行研究にて参考文献は信頼できるものだと定義しているが，誤記述等の内容不一致の可能性を考慮した上で信頼度を算出すべきである．

表2:5件の不一致参照情報の詳細

記事タイトル	情報格差	情報格差	Googleドキュメント	国際交流基金	フェミニズム
カテゴリ	情報	情報	情報	芸術	芸術
資料発表年	2007	2007	不明	不明	2005
ドメイン	ne.jp	ne.jp	com	go.jp	or.jp
著者(組織)	社会実情データ図鑑	社会実情データ図鑑	Google	国際交流基金	日本共産党
リンクの有無	有	有	有	有	有
リンク切れの有無	無	無	無	無	無
記事主張の同異	無	無	無	無	無
記載ページ指定	無	無	無	無	無
情報形態	a	a	a	a	a
Wikipedia記事の整合性	数値の不一致	数値の不一致	数値の不一致	数値の不一致	数値の不一致

次に，これらの不一致参照情報を含んでいた記事そのものについて調査した結果を表3に示す．また比較のために，同じ調査項目について各カテゴリ全体の平均値を表4に示す．

先ほどの表3と表4を比較すると，1000字あたりの参照情報数が不一致箇所を含む記事では平均0.97であるのに対し全記事平均では1.50となり，誤った情報を含んでいた記事は1000字あたりの参照情報数が劣っているとわかる．またリンク切れの割合[5]に関しては，表3と表4を比較してみると，全てのカテゴリにおいて，間違い情報を含む記事のほうが，割合が高いとわかる．

このことから，不一致情報平均は1000字あたりの参照情報数やリンク切れ割合が優れないため，記事の質が低いと判断できる．この2点は不一致情報を含む記事と一般記事を区別することができる指標であるため，信頼性の判断材料になる可能性が高い．

表3:記事別全参照情報の詳細

記事タイトル	情報格差	Googleドキュメント	国際交流基金	フェミニズム	不一致情報平均
カテゴリ	情報	情報	芸術	芸術	
平均資料発表年	2005	不明	2013	2012	2010
1000字あたりの参照情報数	1.28	0.50	1.4	0.68	0.97
参照情報数	26	1	5	9	10.25
リンク設定数	25	1	5	9	10.00
書籍や雑誌数(紙出版物)	0	0	0	0	0
リンク切れ割合	24%	0%	40%	11%	19%

表4:カテゴリ別参照情報の平均値

カテゴリ	情報	医療	歴史	芸術	全記事平均
平均資料発表年	2005	2007	1989	2007	2002
1000字あたりの参照情報数	1.63	1.84	1.47	1.09	1.50
1記事あたりの平均参照情報数	6.67	5.17	9.60	4.10	6.38
1記事あたりの平均リンク設定数	5.10	3.97	3.10	2.20	3.59
1記事あたりの平均書籍数	1.57	1.20	6.50	1.90	2.79
リンク切れ割合	17%	22%	11%	9%	15%

次に，Web情報のドメイン種別の割合をカテゴリごとに算出し，カテゴリごとの傾向について考えてみる．ここでは，先に述べた井上らが信頼できるとしたドメインに「or.jp」と「lg.jp」を追加した5種類のドメインに注目し，参照情報に占める割合を求めた．その結果を図1に示す．

図1から，新聞やニュース情報が多い歴史や政府機関系ドメインが少ないカテゴリがあるなど，カテゴリごとに参照される情報源が異なるため，井上らが信頼できるとしたドメインの割合に違いがあることがわかる．そのため信頼できるドメインが少ないと，カテゴリによっては元々信頼できるドメインの情報が少ないために，極端に信頼性が低いと判断されてしまう可能性がある．これを回避するために，財団法人や社団法人等の非営利法人組織の「or.jp」や地方公共団体の「lg.jp」のような信頼できそうなドメインを少しでも加えたほうが，このような問題にも対処して，より正しい信頼性評価ができる可能性が高まる．図1を見ると，新たに信頼できるドメインを追加したことで全カテゴリにおいて約10%信頼できるドメインが増えたことから，これらを追加したことにより信頼できる情報不足問題を解決できる可能性がある．

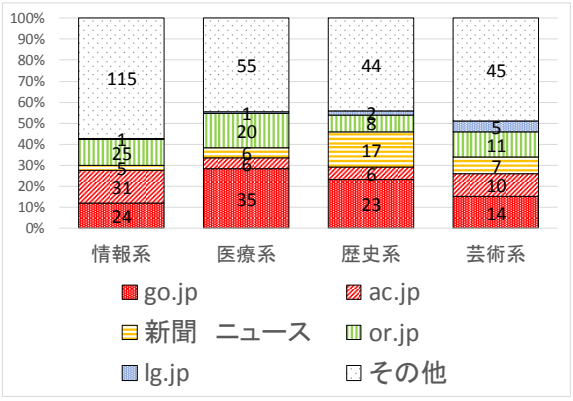

図 1:カテゴリ別の信頼ドメインの割合

図2は各カテゴリの各30記事に含まれている参照情報を形態別に分類したものである．凡例にあるa～dは表1の情報形態と対応している．その結果，「歴史」は紙媒体の出版物が約80%を占めている一方で「情報」は電子データが約90%を占めている．井上らは信頼度算出手法として電子データ情報よりも書籍のほうが信頼できるとし，重み付けを行っていた．

しかし，図2からわかるようにカテゴリによって参照情報形態に大きく差があり，井上らの算出方法では極端に歴史の信頼度が高いと判断されてしまう．そこで，書籍と電子データの信頼度を同等とすることで，情報形態に影響されない正しい信頼度を算出できる．

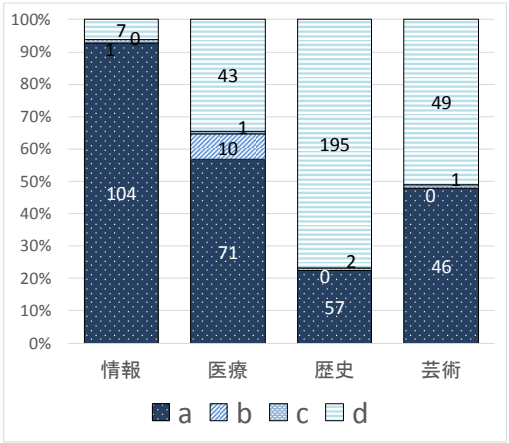

図 2:カテゴリ別情報形態

5 記事ごとにリンク切れ数 / URL 数 となるリンク切れ割合を算出し，それをカテゴリ内のすべての記事で平均化したもの．

ここで対象とするカテゴリは「医療」と「歴史」の2カテゴリである.「秀逸な記事[6]」はWikipediaが定義し認められた記事を医療から5記事, 歴史から5記事を用いる. また「その他」の記事は, 参照情報調査で収集した医療と歴史の記事を用いる.

表5が「秀逸な記事」を対象に各項目の平均値をまとめたものである. 表4, 表5を比較すると, 一記事あたりの平均参照情報数が「秀逸な記事」は多いが, これは内容が多いために「その他の記事」より参照情報数多いことは当然である. 一方で1000字あたりの平均参照情報数は大きな差が見られないことがわかる. またリンク切れ割合に関しては「秀逸な記事」のほうが, 割合が多いため「その他の記事」よりもWikipedia記事内容が参照情報から反映されているかの確認が難しい.

このことから,「秀逸な記事」はWikipediaの定める基準を満たすにもかかわらず, 記事の質が「その他の記事」と同等であるといえる.

表5:カテゴリ別「秀逸な記事」の参照情報平均値

カテゴリ	医療(秀逸)	歴史(秀逸)
平均資料発表年	2003	2000
1000字あたりの平均参照情報数	1.41	1.37
1記事あたりの平均参照情報数	56.0	61.6
1記事あたりのリンク設定数	24.0	27.6
1記事あたりの書籍数	32	34
リンク切れ割合	28%	30%

5. むすび

本研究では, 4カテゴリから各30記事を抽出し, その記事に含まれる参照情報を9の項目に分け調査した. これらの情報を元にいくつかの考察をした.

その結果, 一般記事と間違い情報を含む記事の間に差が見られた1000字あたりの参照情報数やリンク切れ割合, 新たに「or.jp」,「lg.jp」を信頼ドメインに含めること, 電子データと書籍の信頼度を同等にすることで信頼度算出の改善を図れる可能性がある.

今後の課題として, 本研究ではリンク切れや書籍内容の誤記述を考慮した手法を提示したが, 実際に評価式を作成し信頼度算出を行うことができなかった. そのため, 本研究で考えた評価指標を実際に信頼度評価式に用い, 井上らの評価式と比較することが今後の課題として挙げられる.

6 Wikipedia が定める基準を満たしたもので, 百科事典の記事として質・量・書式に問題がない記事のことをさす.

参考文献

[1] Wikipedia, https://ja.wikipedia.org

[2] 長塚隆, 神野こずえ, “学生におけるWikipedia日本語版の利用動向”, 情報知識学会誌, 21, 2, pp.149-156. 2011.

[3] 加藤眞三, WELQ問題「医師監修」だから安全とは限らない, 東洋経済ONLINE, 2016年12月19日, http://toyokeizai.net/articles/-/149965

[4] 金子寛人, WELQ問題, 専門家に聞く「コピペに加筆しても著作権侵害の可能性」, 日経トレンディネット, 2016年12月26日, http://trendy.nikkeibp.co.jp/atcl/pickup/15/1003590/121900702/

[5] 井上雄介, 太田学, “脚注と参考文献を用いたWikipedia記事の信頼性評価の一手法”, 第8回日本データベース学会年次大会, 2010-B10-5, 2010.

[6] 吉次優, 新妻弘崇, 太田学, “参照情報を利用したWikipedia記事の信頼性評価の一手法”, 第13回日本データベース学会年次大会, 2015-D4-2, 2015.

[7] 特許 US6285999, Method for node ranking in a linked database, Google特許検索, https://www.google.com/patents/US62859999

[8] 三上洋, DeNA「WELQ(ウェルク)」休止・・・まとめサイトの問題点と背景は, YOMIURI ONLINE, 2016年12月13日, http://www.yomiuri.co.jp/science/goshinjyutsu/20161212-OYT8T50096.html

[9] 樫原真知子, 武宗次郎, 遠藤有美江, 土井亮平, “Wikipediaの評価”, 慶應義塾大学文学部図書館・情報学専攻上田修一研究会 2007年度グループ研究レポート, 2008.

[10] 日下九八, “ウィキペディア:その信頼性と社会的役割”, 情報管理, Vol.55, 1 , pp.2-12 , 2012.

[11] 近藤敏志, “(91回)キュレーション(知っておきたいワード)“, 映像情報メディア学会誌, 67, pp.695-696, 2013.

[12] 日本経済新聞, キュレーション 王者グーグルを追う人力の新興勢力, 2010年12月30日, http://www.nikkei.com/article/DGXZZO20771920Z21C10A2000000/