文献検索サイトの著者レコードの曖昧性解消における著者との関係の近さと信頼性

日向野達郎† 増田 英孝†† 山田 剛一†† 清田 陽司††† 大向 一輝†††† 中川 裕志†††††

† 東京電機大学大学院 未来科学研究科 情報メディア学専攻 〒120–8551 東京都足立区千住旭町 5
†† 東京電機大学 未来科学部 情報メディア学科 〒120–8551 東京都足立区千住旭町 5
††† 株式会社ネクスト 〒108–0075 東京都港区港南 2–3–13 品川フロントビル
†††† 国立情報学研究所 〒101-843 東京都千代田区一ツ橋 2-1-2
††††† 東京大学 情報基盤センター 学術情報研究部門 〒113–0033 東京都文京区本郷 7–3–1
E-mail: †higano@cdl.im.dendai.ac.jp, ††{masuda,yamada}@im.dendai.ac.jp, †††kiyota@littel.co.jp, ††††i2k@nii.ac.jp, †††††n3@dl.itc.u-tokyo.ac.jp

あらまし 文献検索サービスでは，著者に着目して文献を検索したいというニーズに応えるため，文献レコードから機械的に生成された著者情報に関するレコードを整備している．しかし，同一人物のレコードがスプリットしてしまう問題があり，人手による名寄せが必要とされているが，著者の研究分野に関する知識が必要とされるためクラウドソーシングの適用が難しい．そこで，著者と関係が近い人物にレコードの分類を SNS を介して依頼することで高い信頼性を確保するという手法を提案する．今回，学内の複数の教員を対象として，関係の近さがそれぞれ異なる学生達にレコードの分類を行ってもらうという実験を行なった．実験の結果から，近い関係であるほど正解率は高いということを実証することができた．

キーワード メタデータの統合，MediaWiki，名寄せ，情報修正，クラウドソーシング

1. はじめに

現在 Web 上には，文献検索サイトと呼ばれるポータルサイトが複数存在する．これらのサイトは別々の機関によって運営されているため，サイト間の同一情報のメタデータ同士が対応付けられおらず，複数のサイト上の文献を横断的に検索することができない．そのため網羅的に文献を探しているユーザにとっては，それぞれのサイトで検索を繰り返す必要があるため，非常に手間がかかってしまうというのが現状である．

本研究では，機関の枠を超えて文献情報を横断的に検索することを可能にするシステムの開発を目的としている．このためには各機関がメタデータに対して，それぞれ割り当てている固有の識別番号 (論文 ID，著者 ID 等) を互いに対応付ける必要がある．まず手始めに，著者情報を対象として，MediaWiki 上で複数の文献検索サイトのメタデータを人手で対応付けるための枠組みを提案し，実装している．

人手で膨大な情報を対応付けるためには多くの Wiki 編集者による協力が必要である．そこで編集に参加する動機をもってもらうために，Facebook における友達関係を利用して，編集者と友達関係にあり，Wiki 上で対応付けの編集を必要としている人物を編集者に知らせるという編集支援の機能を提案する．

今回，実際に関係が近いほど正解率が向上するのかどうかを検証するために，東京電機大学内の複数の教員を対象として，関係の近さがそれぞれ異なる学生達にレコードの分類を行ってもらうという実験を行なったので結果を報告する．

2. 文献検索サイトの統合

国立情報学研究所の「CiNii」や，科学技術振興機構の「J-GLOBAL」等の文献検索サービスが様々な機関から提供されているが，横断検索等のサービスの統合はなされていない．例として「CiNii」の著者情報のページから他機関のサービスである「J-GLOBAL」へのリンクというものが存在しているが，「J-GLOBAL」でその著者の名前を検索した結果のページへのリンクであり，直接「J-GLOBAL」の著者情報ページへリンクされているわけではない．これは国立情報学研究所と科学技術振興機構がそれぞれ所有している情報に対して独自に割り当てている識別番号（論文 ID や著者 ID 等）が，互いに対応付けられていないために起こる問題である．

さらに著者データベースでは，著者の所属の異動などによって発生する重複レコードや，同姓同名の複数の人物のレコードを機械的に分類することは難しい [1]．そこで人手による修正が必要になってくるがここでもいくつかの問題がある．情報に間違いがあった場合に機関へ報告しても，機関のスタッフにより確認が行われた後，修正が行われるというように，間違いの発見から修正までに時間がかかってしまう．また他のサイトにおいては，自分でない他の著者の情報に間違いを発見しても修正することができないので，著者本人が間違いに気づくまで情報が間違ったままとなってしまうのが現状である．

そこで本研究では，各機関を間接的につなぎ機関の情報を横断的に検索するサービスの開発を目的とする．さらに，Medi-

aWiki を利用することでアカウントを作成すれば誰でも編集に参加できるので，ユーザが情報に間違いを発見した場合，そのユーザの手で即座に編集をすることも可能になるということが本研究の特徴的な点である．

この目的のために，まず各機関がそれぞれ所有している情報に対して割り当てている ID の対応付けを行い， ID の対応表として利用するためのシステム (機関横断型文献情報 Wiki) を MediaWiki を利用して構築した．

本研究では，まず著者情報を対象として MediaWiki 上で対応付けを行う．図 1 に ID 統合の概念図を示す．

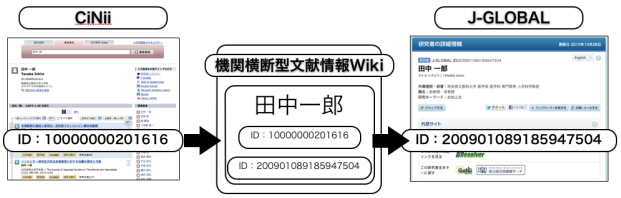

図 1　ID 統合の概念図

図 1 に示すように，従来の文献検索サイトは「CiNii」の「田中一郎」の著者情報ページから，「J-GLOBAL」 の「田中一郎」の著者情報ページへは直接リンクすることができなかった．本システムで各サイト間の同一情報同士の ID を対応付けることで，各サイトを間接的につなぐことを可能にしている．

3.　関 連 研 究

MediaWiki を利用した Wiki システムである Wikipdia と同じように対応付けの編集は，人手によって無償で行われるため，マイクロボランティアのコンセプトを利用する．マイクロボランティアとは，短時間で参加可能なインターネット上のボランティアであり，多数の人の知や力の利用によって問題解決を行うクラウドソーシングの一種である．マイクロボランティアにおいて重要なのは，ひとりひとりの労力を少なくすることで，多くの人に参加してもらうという点である．マイクロボランティアの具体的な例として，2011 年 3 月に発生した東日本大震災直後に 提供された臨時サービス「GooglePersonFinder」[2] が挙げられる．このサービスは，人の安否を知りたいユーザ が相手を特定できる情報を登録しておき，その人物の消息を知る人が現在の状況を投稿することで，消息を伝 えることができる仕組みとなっている．臨時で提供されたにもかかわらず，当時 67 万件以上の安否情報が登録された．マイクロボランティアにおいて重要なことは，ひとりひとりの労力を少なくすることで，多くの人に参加してもらうという点である．

4.　機関横断型文献情報 Wiki

4.1　メタデータの収集

本システムでは，各サイトの所有するメタデータを，サイト内の人物情報ページをスクレイピングすることで機械的に収集する．

各サイトから，メタデータに割り当てられている「ID」と，「名前」,「所属機関」といった人物に関する基本的な情報に加え,「研究分野」や,「論文の一覧」等のように人物を特定するために参考となる情報を収集する．今回は検証のために東京電機大学の教員計 335 名の人物情報を対象とした．

4.2　Wiki への登録

収集したメタデータの Wiki への登録は編集 bot により機械的に行う．ページタイトルはそのサイトにおける ID とする．登録された人物情報の例を図 2 に示す．

図 2　登録されたメタデータページの例

図 2 は著者「増田英孝」の「J-GLOBAL」でのメタデータページである．「J-GLOBAL」における「増田英孝」の ID は「200901009424739052」なのでページタイトルは「J-GLOBAL:200901009424739052」となる．ページ内の項目は，サイトから収集したその人物の基本情報を記載する．このようにしてページを作成していき，同名の著者のページをその人物名をページタイトルとした「人物名ページ」に一覧としてまとめる．

4.3　各サイトのメタデータの対応付け

各サイトから収集したメタデータが Wiki に登録された段階では，各サイト間の対応付けがなされていないので，同一著者のページを対応付けるという作業を行う必要がある．現状ではこの作業は人手で行っている．

編集者は，人物名ページにまとめられたメタデータページ内の，論文の一覧や，所属，研究分野，研究キーワード等の情報をもとに同一人物であるか否かの判断を行い，同一人物であった場合にはページ同士をリダイレクト関係にする．リダイレクト先は最初に Wiki に登録されたページとする．このリダイレクト先のページに対して「リダイレクトしているページのタイトルの一覧」を取り出す．タイトルは各機関における ID なので，このリダイレクト関係が ID の対応表として機能する．

5.　編集支援機能の提案

機関横断型文献情報 Wiki の膨大な情報を人手で対応付けるためには，多くのユーザに編集に参加してもらう必要がある．そこでマイクロボランティアのコンセプトである一人ひとりの負担をできるだけ少なくすることに加え，編集に参加するための動機を持ってもらうことが重要な課題である．

2009 年に行われた Wikimedia 財団による Wikipedia 参加者へのアンケート [3] に「書き込みをしない人が 書き込みをしてくれるようにするためにはどうすればよいと思うか」という問いかけがある．この回答として 「自分の書き込みが必要な特定のトピックがあれば書き込む」というものがあった．このことから，本 Wiki システムの編集に参加してもらうためには，「この編集を行うためには自分の知識が必要である」という意識を持ってもらうということが重要であると考えられる．この点を踏まえて，多くの人々に編集に参加してもらうための編集支援の機能を提案する．

5.1 Facebook の友達関係を利用した編集支援

複数文献検索サイト間のメタデータ対応付けや，重複レコード解消の編集を行うためには，研究内容や，所属などの情報を元に同一人物であるかどうかの判定を行うが，所属情報が記載されていないなど人物を特定するための情報が極端に少ないメタデータも存在する．編集者の専門分野から離れている分野の人物で所属などの情報がなかった場合，同一人物であるか判断することは難しく，間違った編集をしてしまう可能性も高い．そのため，一般的なクラウドソーシングの適用が難しい．

このようなケースに対応するため，Facebook の友達関係を利用して，編集を支援する機能を用意する．図 3 に機能の概要図を示す．

図 3 に示すように，以下の様な流れで処理される機能である．

（1） 編集者用のページを用意し，編集者の Facebook アカウントで OAuth 認証を行なう

（2） Facebook から「編集者と Facebook 上で友達関係にある人物」の名前や職歴、学歴といった情報を取得する

（3） 取得した情報と Wiki のメタデータページの情報で機械的にマッチングを行う

（4） 一致したメタデータページを含む人物名ページの情報を編集者用ページに示す

（5） 編集者に，友達の情報と Wiki のメタデータページの人物情報を比較し，同一人物であるかを判断してもらう

という機能である．

Facebook で友達関係にあるということは，編集者は対応付けが必要な人物についてある程度詳しいと考えられるため，対応付けの編集も比較的容易であり，編集の信頼性が向上すると考えられる．

6. 関係の近さと信頼性に関する実験

Facebook の友達関係を利用した編集支援では，著者と関係が近い人物ほど信頼性の高いレコードの分類を行うことが可能であるという仮説を元にしている．この仮説が正しいかどうかを検証するために関係の近さと信頼性に関する実験を行った．

6.1 実 験 方 法

東京電機大学の教員の中から 4 名を実験対象教員として選定する．被験者（東京電機大学の学生 12 名に，実験対象教員の名前と一致する CiNii の著者レコードそれぞれ 10 件を，実験対象教員の J-GLOBAL の著者レコードと同一人物であるかどうか判断してもらった．それぞれのレコードに対して「同一人物」「別人物」「わからない」という判断に加え，その判断に「自信がある」か「自信がない」かについても回答してもらった．正しく判断できた回答を判別成功とし，関係の近さと回答の判別成功率の関係を調べる．仮説では，被験者と関係が近い教員ほど回答の正解率が高くなると考えられる．

6.2 関 係 性

今回の実験では，教員と学生の関係の近さを計るため以下の問に答えてもらう．ここでは関係の近さを関係性と定義している．

- 問 1．この人物を知っている
- 問 2．この人物の授業を履修したことがある
- 問 3．この人物の研究内容を知っている
- 問 4．この人物の研究指導を受けている
- （問 4 が「はい」なら）問 5．大学生か大学院生か

これらの問に対して「はい」（または「大学院生」）ならば関係性+1,「いいえ」（または「大学生」）ならば関係性+0 とすることで，0〜5 の 6 段階の関係性を量る．関係性が 5 に近いほど関係が近いということになる．

6.3 レコードの判断難易度

著者レコードにおいては主に「所属」「研究内容」「共著者」が人物を特定するための有用な情報となる．しかし重複レコードにおいてはいずれかの情報が欠けていることが多く，レコードごとに情報量が異なる．問題に使用する 10 件のレコードを選定する際に無作為に抽出した場合，実験対象教員ごとに情報量の合計が異なるので，関係の近さに関わらず正解率に差がでてしまう．そこで情報量をある程度均一にするために，情報量という観点から，目的の人物と同一人物に関するレコードかどうか判断する際の難易度を以下の 3 つのグループに分類した．

- 難易度 1「情報量が多く，関係性が 0〜1 でも判別可」
- 難易度 2「情報量が不十分，関係性が 2〜3 なら判別可」
- 難易度 3「情報量が少なく，関係性が 4〜5 なら判別可」

3 つのグループから実験対象教員ごとに 3:4:3 の割合でレコードを抽出し 10 件のレコードを選定することで，実験対象教員ごとの情報量をある程度均一にすることができる．

6.4 実 験 結 果

関係性ごとの回答の判別成功率の平均をまとめた表を表 1 に示す．上段は，「自信がない」回答も正解していれば含めた場合の結果，下段は「自信がない」回答を除外した場合の結果となっている．表 1 の結果から，関係性が高い（関係が近い）ほど判別成功率が高いという結果が得られた．「分からない」回答を除き，判別ができた回答の内「自信がある」回答と「自信がない」回答の正解率の平均を表 2 に示す．

表 1　関係性ごとの回答の判別成功率の平均

	関係性					
	0	1	2	3	4	5
自信がない含む	61 %	55 %	74 %	91 %	93 %	85 %
自信がない除く	32 %	35 %	52 %	68 %	68 %	74 %

表 2 より，自信がある回答は関係性に関わらず正解率が高く，

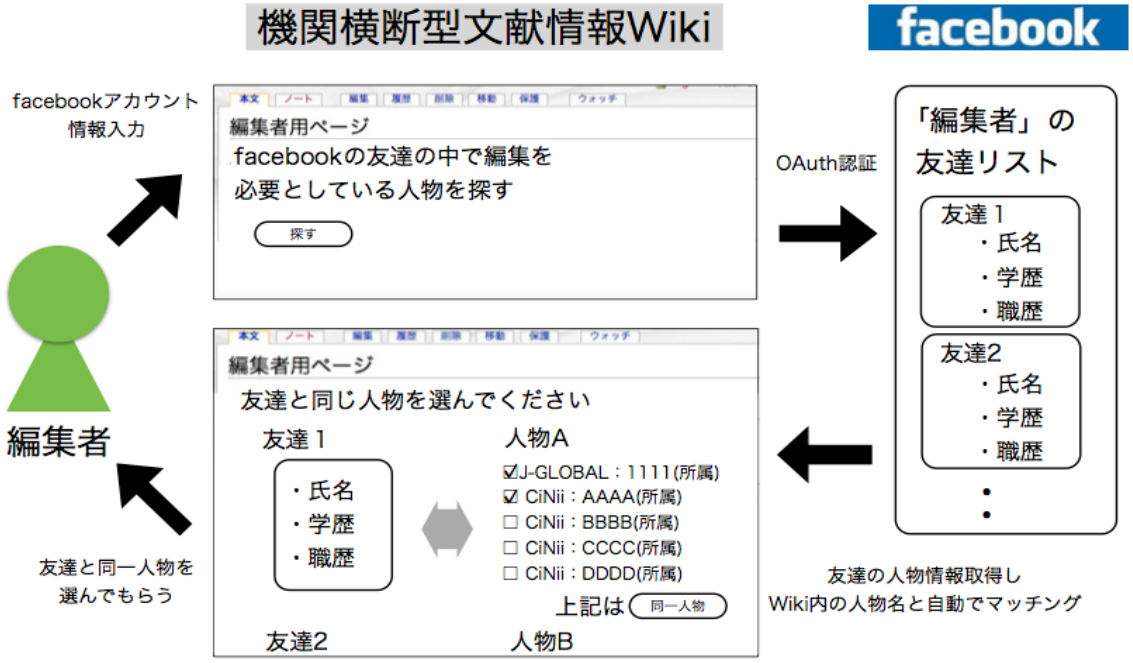

図 3　Facebook の友達関係を利用した編集支援機能の概要

表 2　関係性ごとの回答の正解率の平均

	関係性					
	0	1	2	3	4	5
自信がある	94.9 %	93.8 %	98.3 %	100 %	100 %	92.2 %
自信がない	84.8 %	70.0 %	96.0 %	100 %	100 %	90.0 %

自信がない回答でも関係性が高くなるほど正解率が高くなることが確認できた．

7. 考　　察

本システムと同様に MediaWiki を利用したサービスであるオンライン百科事典「Wikipedia」でも問題視されている通り，サービス内の情報の信頼性という点では，多くの課題がある．アカウントさえあれば誰でも自由に編集できるという MediaWiki の特徴から，誤った編集を行ってしまったり，悪意ある編集者がでたらめな情報を追加するという可能性がある．しかし間違った編集が行われていることに他のユーザが気づけば，そのユーザの手で元の状態に戻すことが可能である．そのため多くのユーザが利用してくれるようになれば，情報の修正も多く行われるようになるため，ある程度の情報の信頼性は確保できるものと考えられる．

関係の近さと信頼性に関する実験結果から，教員と学生間の関係においては，関係が近くなるほど信頼性のある編集を行うことができるということを確認できた．そして正解率の結果から自信がない回答でも関係が近ければ正解率は高いということが確認できた．しかし，Wiki で編集を行う場合，ユーザが自信がない情報を編集するということは期待できない．そのため自信がない場合でも編集に参加しやすくするための支援機能を考える必要があると思われる．

8. お わ り に

既存の文献検索サイトは，それぞれが別々の機関により提供されているので，メタデータの対応付けがなされておらず，横断的に文献を探しているユーザはそれぞれのサイトで検索を繰り返す必要があった．

そこで我々は，各機関が情報に対して割り当てている固有 ID の対応付けを行い各サイトの情報を間接的につなぐための仕組みである機関横断型の文献情報統合システムを MediaWiki を利用することで構築した．このシステムを ID の対応表として利用することで，機関を横断して情報を収集するということが可能になる．

そして多くの人びとに編集に参加する動機をもってもらうために，Facebook における友達関係を利用して，関係の近い著者を編集者に知らせるという編集支援の機能を提案した．この機能では，関係が近いほど信頼性の高い編集ができるという仮説に基づいている．仮説が正しいかを検証するために実験を行ない，関係が近いほど判別の精度が高いという傾向があるということを確認できた．

自分の判断に自信がない場合でも編集に参加できるようにするための支援機能を考案することが今後の課題である．

文　　献

[1] 相澤 彰子 他, レコード同定問題に関する研究の課題と現状, 電子情報通信学会論文誌 J88-D-I, No.3, pp.576–589 (2005).
[2] Google Person Finder, http://google.org/personfinder/global/
[3] R. Glott, P. Schmidt, R. Ghosh, ”Wikipedia Survey ? First Results”, http://upload.wikimedia.org/wikipedia/foundation/a/a7/Wikipedia General Survey-Overview 0.3.9.pdf