E-059

Wikipedia 記事に対する類似記事群の出典傾向の提示方式

Representation of the citation tendency of the similar article group in the Wikipedia

北村 大樹 †　　山田 剛一 †　　絹川 博之 †

Hiroki Kitamura　Koichi Yamada　Hiroshi Kinukawa

１　はじめに

Wikipedia では情報の検証可能性を満たすため，出典を明記することを要求している．しかし，出典が付けられている記事は少数にとどまっていることが調査によりわかった[1].

Wikipedia 記事の信頼性を高めるには，必要な出典を付与することが不可欠である．しかし，それを完全に人手で行うのは作業負担が大きい．そこで，Wikipedia 記事の編集者が，新規に作成した記事，あるいはすでに存在している記事に出典を付与する際に，付けるべき出典の媒体(例: Web ニュース，書籍)を提示するシステムを構築したい．

我々は，出典・脚注の付けられている Wikipedia 日本語版の記事に関して，出典媒体の使用傾向(以下，出典傾向とする)を調査している．今回は，出典情報を付与すべき記事と出典傾向が類似すると考えられる記事グループを求め，その出典傾向に基づいて編集者に付与すべき出典の媒体を提示するシステムを提案する．記事グループは，記事に付与されている各カテゴリに属する記事集合の演算によって定められる

２　出典・脚注情報と記事の出典傾向

2. 1 出典・脚注情報の現状

Wikipedia の内部表現において，出典・脚注情報は <ref> というタグを用いて本文中に示されている．

Wikipedia 日本語版には，2011年5月22日時点で 188,894 件の記事に 1,032,958 件の出典・脚注情報(ISBN 付き参考文献情報含む)が存在する．

2. 2 出典傾向

<ref> タグをもとに抽出した情報は出典と脚注が混在している． Wikipedia 記事の出典傾向を調査するためには，これらを出典と脚注に分ける必要がある．以前報告した出典媒体分類器[1](精度 96.3 %)を用いて出典媒体を分類し，それぞれの使用割合を求めた．

今回は以下の分類を使用した．

(1) 書籍	(2) 論文	(3) 新聞
(4) 雑誌･学会誌	(5) Web：政治系	(6) Web：スポーツ
(7) Web：ブログ	(8) Web：その他	(9) ラジオ番組
(10) TV番組	(11) その他	(12) 脚注

このうち，「脚注」は出典情報でないので，出典数の合計値に含まない．また，「Web：政治系」「Web：ニュース」「Web：スポーツ」「Web：ブログ」については，人手により作成した URL リストを用いて，それぞれ判断している．

† 東京電機大学大学院未来科学研究科
Graduate School of Science and Technology
for Future Life, Tokyo Denki University

2. 3 出典傾向の類似する Wikipedia 記事の出典傾向調査

同一のカテゴリが付与されている記事群を取り出し，それらの出典傾向を調査した．単独のカテゴリで取り出した記事群の出典傾向の分散が大きい場合は，複数のカテゴリが共通して付与されている記事群をとる．

今回は，人物に関する記事について，その記事に付与されているカテゴリを持つ記事群の出典傾向を取り出し，カテゴリ「存命人物」を持つか持たないかによりグループを分けて調査した結果を掲載する．

調査結果の一例として，スポーツ選手の代表「イチロー」，政治家の代表「菅直人」，それぞれのグループの結果を掲載する．「イチロー」ではカテゴリ「MLBオールスターゲーム選出選手」を持つ記事のうち，カテゴリ「存命人物」を持つ記事グループを表１，持たない記事グループを表２に示す．「菅直人」ではカテゴリ「衆議院議員」を持つ記事グループのうち，「存命人物」を持つ記事グループを表３，持たない記事グループを表４に示す．

なお，４つの表からは，共通して値が 0 しか存在しない「論文」，「 Web：ブログ」，「ラジオ番組」の列を除いている．行は出典付与合計数によりソートした．また，行中最も値の高いセルを，背景反転により強調している．

出典傾向の似た記事集合を取得する手法としては，扱う事象の類似する Wikipedia 記事をグループ化することが考えられる．記事の所属する概念を一意に決定し，その概念ごとにまとめるやり方として， Wikipedia のカテゴリ構造を利用した手法[2]が提案されている．この手法を用いて Wikipedia 記事をグループ化した場合，少数の上位概念によって記事群がまとめられるため，記事群の出典傾向が一定にならない．目標とする出典傾向の類似する記事のグループ化には，より狭いまとまりを作る必要があるため，この手法は適用できない．

３ 考察

3. 1 出典傾向について

存命の人物はスポーツ選手・政治家ともに Web 出典がよく用いられていることがわかった．特に，カテゴリ「存命人物」が付けられた政治家の記事群の場合， Web 政治系がよく参照されていた．スポーツ選手の記事群の場合，政治家ほど明確な傾向は見られないものの， Web サイトが比較的多い． Web スポーツ系の値が大きくないのは，URL リストの登録漏れによるものだと考えられる．また，カテゴリ「存命人物」が付けられていない記事群では，書籍が比較的多く用いられている傾向が見られた．

カテゴリ「存命人物」が付けられた記事で Web 出典が多用されていることについては，更新頻度の高い Web 出典が，現在進行形の事象の出典として用いられやすいこと

表１ カテゴリ「MLBオールスターゲーム選出選手」と「存命人物」を持つ Wikipedia 記事の出典傾向

	書籍	新聞	雑誌･学会誌	Web：政治系	Web：スポーツ	Web：その他	TV番組	その他	合計出典数
松井秀喜	0.01	0.12	0.02	0.33	0.12	0.36	0.01	0.02	335
イチロー	0.01	0.02	0	0.27	0.26	0.35	0.01	0.08	136
デレク･ジーター	0.02	0	0.07	0.26	0.19	0.42	0.02	0.02	57
ロイ・ハラデイ	0.07	0.05	0.10	0.21	0.24	0.33	0	0	42
トレバー・ホフマン	0.13	0	0.03	0.29	0.26	0.29	0	0	38
ライアン・ハワード	0.16	0.03	0.03	0.22	0.27	0.30	0	0	37
アルバート・プホルス	0.17	0	0.17	0.19	0.22	0.25	0	0	36
トロイ・トゥロウィツキー	0.08	0	0.14	0.14	0.36	0.28	0	0	36

表２ カテゴリ「MLBオールスターゲーム選出選手」を持ち「存命人物」を持たない Wikipedia 記事の出典傾向

	書籍	新聞	雑誌･学会誌	Web：政治系	Web：スポーツ	Web：その他	TV番組	その他	合計出典数
ジャッキー・ロビンソン	0.71	0	0.03	0.03	0.10	0.08	0	0	62
ロベルト･クレメンテ	0.64	0	0.02	0.09	0.18	0.07	0	0	45
ジョー･ディマジオ	0.13	0	0.13	0.13	0	0.63	0	0	8
ケン・ボイヤー	0	0	0	0	0	1.00	0	0	6
マーク･フィドリッチ	0	0	0.17	0.17	0.33	0.33	0	0	6
サチェル・ペイジ	0.80	0	0	0.20	0	0	0	0	5
ジャッキー・ジェンセン	0.60	0	0	0.20	0	0.20	0	0	5
ボブ・フェラー	0	0	0	0.25	0.25	0.50	0	0	4

表３ カテゴリ「衆議院議員」「存命人物」を持つ Wikipedia 記事の出典傾向

	書籍	新聞	雑誌･学会誌	Web：政治系	Web：スポーツ	Web：その他	TV番組	その他	合計出典数
鳩山由紀夫	0.06	0.07	0.02	0.62	0.02	0.17	0	0.04	305
森喜朗	0.25	0.09	0.10	0.25	0.01	0.27	0	0.03	233
佐藤ゆかり	0.05	0.03	0.13	0.42	0.08	0.29	0	0	230
福田康夫	0.04	0.20	0.02	0.6	0.01	0.12	0	0	223
麻生太郎	0.08	0.15	0.04	0.33	0.06	0.20	0.01	0.12	202
石原慎太郎	0.32	0.14	0.04	0.25	0.01	0.23	0	0.03	200
安倍晋三	0.13	0.15	0.04	0.48	0.02	0.18	0	0.01	181
小沢一郎	0.14	0.05	0.03	0.5	0.01	0.23	0	0.03	120

表４ カテゴリ「衆議院議員」を持ち「存命人物」を持たない Wikipedia 記事の出典傾向

	書籍	新聞	雑誌･学会誌	Web：政治系	Web：スポーツ	Web：その他	TV番組	その他	合計出典数
中川昭一	0.13	0.11	0.03	0.34	0	0.24	0.02	0.14	101
中井一夫	0.77	0.23	0	0	0	0	0	0	44
中川一郎	0.88	0	0	0.09	0	0.03	0	0	33
三塚博	1.00	0	0	0	0	0	0	0	8
三木武夫	1.00	0	0	0	0	0	0	0	5
中山泰秀	0	0	0	0.25	0	0.75	0	0	4
中川俊思	0.50	0	0	0	0	0	0	0.50	4
上村進	0.67	0	0	0.33	0	0	0	0	3

を示していると考えられる．

3. 2 出典傾向の類似する記事のグループ化について

今回は同一カテゴリを持つ記事グループで，出典傾向が類似しているかを確認した．同じ野球選手，同じ衆議院議員であっても，存命中とそうでない場合で出典傾向が異なることから，出典傾向の類似する記事をまとめるには，より狭い範囲でのグループ化が有効である．今回の人物記事のように，場合によってはカテゴリの論理演算(この場合は論理積)をすることが必要になると考えられる．

システムを実装する際，編集者が複数のカテゴリを付与している場合には，その組み合わせの中から出典傾向が安定しているものを選び，それを元に出典を提示する方式を検討している．

4 出典の信頼度

本研究と同じく，出典・脚注情報を扱う研究に，井上らの研究[3]がある．この研究では，脚注中の出典情報と，参考文献情報を活用し， Wikipedia 記事の信頼性を自動評価するシステムを提案し，評価実験をしている．この研究ではすべての Wikipedia 記事に関して一律の評価基準を用いている．

これに対し我々は，付与される記事の種別によって妥当な出典が異なるという立場をとる．例えば，スポーツに関連する話題であれば，スポーツ新聞やスポーツ関連の Web サイトを出典として用いても構わない．その場合，スポーツ新聞をスポーツ関連の記事に限って「出典として妥当」だと判断する．

5 おわりに

5. 1 得られた成果

出典傾向の似た記事集合を作る際に，カテゴリを組み合わせることで，出典傾向が定まってくることがわかった．この出典傾向に基づき，付与すべき出典媒体の提示が可能になる．

5. 2 今後の課題

Wikipedia 記事のグループ化手法の改良を進める．

また，今回得られたデータを基にした出典提示システムを今後実装する．出典として用いられる媒体自体の信頼性についても取り扱うことを考えている．

参考文献

[1] 北村大樹，山田剛一，絹川博之：『 Wikipedia 出典・脚注情報の媒体分類の自動付与』: 情報科学技術フォーラム講演論文集 9(2), 303-304, 2010-08-20

[2] 白川真澄，中山浩太郎，原隆浩，西尾章治郎：『 Wikipedia のカテゴリネットワークを用いた概念のベクトル化手法』: 情報処理学会研究報告. データベース・システム研究会報告 2008(56), 89-96, 2008-06-12

[3] 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(O)』: DEIM Forum 2010 B10-5