E-037

# Wikipedia 出典・脚注情報の媒体分類の自動付与
# Classification of Citation Information in Japanese 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 日本語版には，およそ 60 万件の出典・脚注情報が存在する．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．一方で，現在の Wikipedia には十分な数の出典がつけられていない記事が多く存在する．

記事の信頼性向上を図るには，必要な出典を付与することによって，こうした状況を改善することが不可欠である．それを人手で行うには作業負担が大きいので，出典を半自動的に検索・推薦・付与するシステムを構築することが望ましい．そのためには，Wikipedia エントリに付与するのにふさわしい出典にはどんなものがあるのか，また，実際にはどのような媒体が出典として用いられているのかを調べる必要がある．

本論文では，出典・脚注情報から自動的に出典情報を分離し，その分類を判定するシステムを提案する．

## 2 出典・脚注情報の抽出

### 2. 1 出典・脚注情報の示され方

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(例外あり)．

### 2. 2 出典・脚注情報の付与状況

Wikipedia 日本語版の全エントリに付与されている出典・脚注情報を，正規表現を用いて取り出した．2010年6月7日現在，137,050 件のエントリに 683,972 件の出典・脚注情報が存在する．情報を取り出した段階では出典と脚注の情報が混じっているため，出典情報だけを参照したい場合，出典と脚注を区別する必要がある．

## 3 出典・脚注情報の自動分類

### 3. 1 出典分類の判定方式

取り出した出典・脚注情報に対して，以下に述べる手順で判定する．この際，いずれかの出典情報だと判断された段階で処理を終了する．付与する出典分類は歴史書，新書，文庫本，新聞，雑誌，漫画雑誌，一般書籍，論文，URL，テレビ局，ラジオ局，上記以外のテンプレートの分類，その他とする．以下，分類手順を説明する．

(1) 出典テンプレートを利用した分類

Wikipedia で使用されている出典テンプレートを用いて示された情報は出典であることが明確である．書籍と URL の分類はまとめ，それ以外の分類はテンプレートに従う．

(2)正規表現による分類

URL や書誌情報の特徴的な表現を取り出す正規表現を用いて，出典・脚注情報の媒体を分類する．

(3) Wikipedia 一覧リスト照合による分類

Wikipedia の一覧記事から作成した「新聞」「雑誌」「文庫本」「新書」「歴史書」「出版社」「テレビ局」「ラジオ局」リストの各要素と照合し，それを含んでいる上でさらに特定表現を持つものを，それぞれの出典情報と判定する．

(4) 特定表現による出典情報の分類

出典・脚注情報の中には，末尾に「本人談」「参照」「引用」「発言」「抜粋」「記載」などといった語をおくものがある．情報源を記述しているこれらを「その他の出典」と判定する．

(5) ナイーブベイズ分類器による分類

(1)から(4)の処理で分類されなかった出典・脚注情報のうち，論文を示す出典情報は表記揺れが非常に大きいため，ナイーブベイズ分類器[1]により判定する．学習データには人手で収集した論文 345 件，書籍 632 件，脚注 716 件の出典・脚注情報を使用している．

### 3. 2 出典・脚注 Web ページの信頼度判定

Web ページは更新頻度が高いので，最新の事情についての出典として用いられやすい傾向が見られる．その質は玉石混合で，信頼性を測る研究が従来よりなされている．今回は，出典・脚注情報中で参照されている Web ページについて，先行研究[2][3]を参考に，信頼性を自動判定するための基準を提案する．

各基準の頭についている記号は，(+) は肯定的な評価を与える，(−) は否定的な評価を与える，(±) は両方の評価を含む基準であることを意味する．

(1) URL を見て判断できるもの
- (+)記事配信主体の社会的評価
  上場企業リストを使用
- (+)ドメイン名
  go や or は高めに評価
- (−)不特定多数ユーザがコンテンツを作成できるか
  URL に "wiki", "bbs", "blog" が含まれているかを確認
- (±) Wikipedia 内でのホストの被参照エントリ件数
  何件のエントリがそのホストを出典・脚注で参照しているかを確認
- (+) 秀逸・良質な記事で参照されているか
  秀逸・良質な記事で参照されている URL ホストが含まれているかを確認

(2) URL の参照先から文書を取得して判断するもの
- (−) Web サイトの寿命
  サイトにアクセスし，レスポンスコードを確認

† 東京電機大学大学院 Tokyo Denki University

(−)登録・ログイン等が必要であるか
　　フォームの有無を確認
(+)著作権・著者情報表示が存在
　　© マーク等の有無を確認

以下，この基準について述べる．(1)は文字列を確認するだけなので，すぐに計算できる．(2)の計算は一つ一つの URL から文書を取得する必要があるため，時間がかかる．現在の Wikipedia 日本語版の出典・脚注中には約 23 万件の URL が存在しているため，より効率の良い手法を採ることが望ましい．

良い出典・脚注 URL の見本が多く存在する一方で，悪い見本が見つからない．これは Wikipedia 内において編集者同士の相互監視が働いていて，不適切な出典が残らないようになっているからだと考えられる．

## 4　実験・評価

今回は 3.1 の処理について実験した．無作為に選択した 300 件のエントリに付与されていた 1,236 件の出典・脚注情報に対して分類を判定した．縮約した結果を表 1 に示す．

### 4. 1 各分類法ごとの評価

テンプレートと URL の分類については，その定義上，完全なルールが用意されているため，100 %正解となる．

それ以外の分類法について評価する．他と比較してナイーブベイズ分類器の精度が低く，書籍で 67.6 %，論文で 71.9 %，全体で 88.4 %となっている．再現率も 90.7 %と，他の方法に比べると低い．他の分類法は精度が 95 %以上，再現率が 90 %以上となっている．

### 4. 2 全体の評価

精度，再現率ともに 96.3　%となっている．

表 1　出典媒体の分類結果

| | | 書籍 | 論文 | URL | その他 | 全体 |
|---|---|---|---|---|---|---|
| テンプレート | H正解 | 54 | 0 | 20 | 80 | 154 |
| | M正解 | 54 | 0 | 20 | 75 | 149 |
| | M不正解 | 0 | 0 | 0 | 0 | 0 |
| | 精度(%) | 100.0 | 0 | 100.0 | 100.0 | 100.0 |
| | 再現率(%) | 100.0 | 0 | 100.0 | 93.8 | 96.8 |
| 正規表現 | H正解 | 192 | | 359 | | 551 |
| | M正解 | 187 | | 357 | | 544 |
| | M不正解 | 2 | | 0 | | 2 |
| | 精度(%) | 98.9 | | 100.0 | | 99.6 |
| | 再現率(%) | 97.4 | | 99.4 | | 98.7 |
| 特定表現 | H正解 | | | | 43 | 43 |
| | M正解 | | | | 39 | 39 |
| | M不正解 | | | | 2 | 2 |
| | 精度(%) | | | | 95.1 | 95.1 |
| | 再現率(%) | | | | 90.7 | 90.7 |

| | | 書籍 | 論文 | URL | その他 | 全体 |
|---|---|---|---|---|---|---|
| リスト照合 | H正解 | 175 | | | 1 | 176 |
| | M正解 | 175 | | | 0 | 175 |
| | M不正解 | 4 | | | 1 | 5 |
| | 精度(%) | 97.8 | | | 0 | 97.2 |
| | 再現率(%) | 100.0 | | | 0 | 99.4 |
| NB分類器 | H正解 | 25 | 46 | | 241 | 312 |
| | M正解 | 25 | 46 | | 212 | 283 |
| | M不正解 | 12 | 18 | | 7 | 37 |
| | 精度(%) | 67.6 | 71.9 | | 96.8 | 88.4 |
| | 再現率(%) | 100.0 | 100.0 | | 88.0 | 90.7 |
| 全体 | H正解 | 446 | 46 | 379 | 365 | 1236 |
| | M正解 | 441 | 46 | 377 | 326 | 1190 |
| | M不正解 | 18 | 18 | 0 | 10 | 46 |
| | 精度(%) | 96.1 | 71.9 | 100.0 | 97.0 | 96.3 |
| | 再現率(%) | 98.9 | 100.0 | 99.5 | 89.3 | 96.3 |

H正解:　人手により判断した正解
M正解:　機械処理結果における正解
M不正解: 機械処理結果における不正解
精度　=　M正解 / (M正解 + M不正解) * 100 [%]
再現率 =　( M正解 / H正解 ) * 100 [%]
書籍は一般書籍，歴史書，新書，文庫本，新聞，雑誌，漫画雑誌を含む
その他はテンプレートの分類[4]のうち書籍，論文，URL に該当しないものと，TV 番組，ラジオ番組を含む

## 5　考察

### 5. 1 URL 出典情報について

今回，URL で示された出典情報については，無条件でWeb 上の出典だと判断した．実際には URL が含まれていても，それが必ずしも Web サイトによる出典であるわけではない．例として，脚注の根拠となる Web サイトを示すために URL を用いている場合がある．これについては，URL を使用している脚注を学習データとして用意し，出典・脚注情報との類似度を計測して，その数値が高い場合には脚注であると判断する解決策が考えられる．

### 5. 2　出典情報の傾向

実験結果の正解を見ると，出典に用いられている媒体は書籍，URL の順に多くなっている．Web ページは編集者・閲覧者共に他媒体と比べ内容をすぐ参照でき，また情報が更新されやすいことから時事問題等のエントリで出典として用いやすいので，多用される傾向がある．

### 5. 3 関連研究と本研究の位置づけ

Wikipedia の出典情報を用いた先行研究には Nielsen の研究[5][6]がある．これは Wikipedia 英語版の科学分野における出典について調査・報告したものである．

対象とする言語版は，Nielsen は英語版であるのに対し，本研究では日本語版であるという違いがある．

出典情報と脚注については，テンプレートを用いた出典情報のみを対象とすることで区別している．この点が，全分野・全エントリの出典・脚注情報を対象とした本研究と最も異なっている．

## 6　おわりに

### 6. 1 得られた成果

Wikipedia の出典・脚注情報を分類判定するシステムを作成した．このシステムは実験において，出典・脚注情報を精度・再現率ともに 96.3　%で分類することができた．

### 6. 2 今後の課題

本論文にて提案した URL の信頼度を判定するシステムを今後実装する．

## 参考文献

[1] Christopher M. Bishop 著,『パターン認識と機械学習 上 - ベイズ理論による統計的予測』シュプリンガー・ジャパン, 2007
[2] ヴェラヤサン ガネサン,　山田誠二『 Web ページの相対信頼度』人工知能学会全国大会論文集, 2004
[3] 福島 隆寛, 内海 彰,『 Web ページの信頼性の自動推定』, 知能と情報, Vol. 19, No. 3. pp.239-249, 2007
[4] 『Category:出典テンプレート』 - Wikipedia 日本語版 http://ja.wikipedia.org/wiki/Category:出典テンプレート
[5] Finn Årup Nielsen , "Scientific Citations in Wikipedia" First Monday, Vol.8, No.12 , 2007
[6] Finn Årup Nielsen , "Clustering of scientific citations in Wikipedia" Arxiv preprint arXiv:0805.1154, 2008