D-035

Web ページの著者の同定

Identification of the Author of Web Pages

加藤 義清† 河原 大輔† 乾 健太郎† 黒橋 禎夫†‡ 柴田 知秀‡
Yoshikiyo Kato Daisuke Kawahara Kentaro Inui Sadao Kurohashi Tomohide Shibata

1．まえがき

いまや Web ページは，ビジネスや日常生活において様々な意思決定をする際の重要な情報源となっている．しかしながら，近年，ブログや動画共有サイトなどに代表されるユーザ発信型コンテンツ (UGC) の急速な普及により，Web にある情報の玉石混淆の度合いは著しくなっており，その信頼性を判断しながら活用することが益々困難となってきている．

現在 Web から必要な情報を見つけ出す手段として，検索エンジンの利用が一般的となっている．しかし，検索エンジン最適化 (SEO)といって，検索結果の中でより上位にランクされるために Web ページに対して様々な工夫を凝らすというようなことが広くおこなわれており，検索結果の上位にランキングされる Web ページに書かれている情報が必ずしも信頼できるわけではなく，信頼できる情報を見つけ出すには困難が伴う．

本研究では，Web 上の情報の信頼性を評価するために，特定のトピックについてある程度まとまった量の情報に対して分析をおこない，その内容について発信者による意見の分布や，主要な言論とそれに対立する言論など，俯瞰的な情報を分析し，個々の情報を全体の中で位置づけることにより信頼性の判断を可能にすると言うアプローチを取る[1]．特に本研究が対象とするのは情報の発信者についての分析である．

著者らはこれまでに，Web ページの情報発信者を，情報発信への役割も含めて捉えるために情報発信構成として記述することを提案している[2]．次節以降では，まず情報発信構成の考え方について概観し，Web ページの著者の同定を，情報発信構成同定の部分問題として位置付ける．その後，Web ページの著者同定について手法を提案し，評価結果について報告する．最後に本研究の今後の方向性について述べる．

2．Web ページの情報発信構成

Web ページの情報発信構成とは，ある Web ページの情報発信者，その情報発信クラス，および情報発信者間の関係を与えるものである．

Web ページの情報発信者とは，Web ページに含まれる情報の内容，およびその公開について責任を有する個人や団体などの実体を意味する．情報発信者には Web ページの著者，Web ページを公開する Web サイトの運営者（サイト運営者），Web ページの中で引用された情報の著者などが含まれる．例えば，図 1 に示した Web ページには 2 つの情報発信者がある．1 つはサイト運営者である「NICT 独立行政法人情報通信研究機構」であり，もう 1 つは Web ページの内容の著者である「宮原秀夫」である．

図 1: Web ページの情報発信者.

情報発信者の情報発信クラスとは，情報発信者を団体・個人，営利・非営利などの軸により「個人・専門家」「企業」「政府機関」「大学」など 15 のクラスを定義したものである．図 1 の例で言えば，「NICT」は「政府機関」に，「宮原秀夫」は「専門家」にそれぞれ分類される．

情報発信構成において，情報発信者間の関係を表すために情報発信タイプが定義される．現在のところ「所属発信者タイプ」「掲載タイプ」など 6 種類のタイプを定義している．図 1 の例ではサイト運営者（NICT）に所属する著者（宮原秀夫）が発信しているので，所属発信者タイプとなる．

情報発信構成はこれまでに述べた情報を下記の記法により表現される．

（＜発信タイプ＞，＜サイト運営者＞，＜著者＞, ...）

更に，サイト運営者や著者は情報発信クラス，名前，肩書き，所属組織の組として表現される．例えば，図 1 の Web ページの情報発信構成は次のように表現される．

（所属発信者,
（政府機関，”情報通信研究機構”）
(-，”宮原秀夫”, “理事長”, -）
）

ここでは情報発信構成の概略について説明した．詳細については文献[1]などを参照されたい．

3．Web ページの著者の同定

情報発信構成の同定問題とは，前節で述べた情報発信構成に含まれる全ての情報を同定する問題である．このうち，サイト運営者を同定する手法については既に報告済みである[2]．本稿では，Web ページの著者を同定するための手法を報告する．

著者同定処理の流れを図 2 に示す．まず，処理対象の Web ページから著者名の候補となる名前を抽出する．その後，主要コンテンツからの距離など，各候補に与えられる

†（独）情報通信研究機構, NICT
‡京都大学, Kyoto University

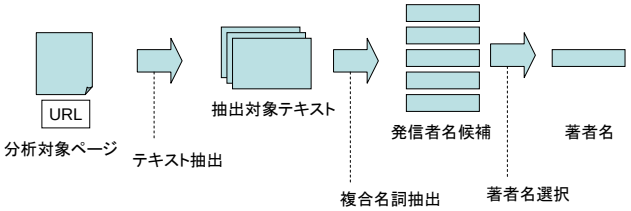

図 2: Web ページの著者同定処理の流れ

素性に基づきランキングをおこない，ランキングが 1 位の候補をそのページの著者として出力する．ここでは，Web ページ内に著者が一人しかいない場合のみを扱うことにする．

3.1 著者名候補の抽出

著者名候補の抽出は以下の手順でおこなわれる．

（1） HTML からテキストを抽出する
（2） テキストを文単位に分割する
（3） 以下の条件を満たす文のみを残す
(ア) 「の」以外の助詞の文の総形態素数に対する割合がある閾値（実験では 0.1）以下であること
(イ) 連用の複合辞（「～について」など）を含まないこと
（4） 残った文の各文節から複合名詞を抽出する
（5） 複合名詞のうち，以下の条件を満たすものを著者名の候補として残す．
(ア) 品詞分類が「人名」「組織名」である形態素を含む
(イ) 語末の形態素が「人名末尾」「組織名末尾」である
(ウ) 未知語を含む

3.2 主要コンテンツからの距離

著者名は多くの場合，主要コンテンツの冒頭あるいは末尾に記されていることが多い．この性質を利用するために，著者らしさを判別するための素性として，文書構造上での著者名候補と主要コンテンツとの距離を示す尺度を導入する．

距離尺度を導入する前段階として HTML の DOM の平坦化をおこなう．DOM は HTML の文書構造を木構造として表現したモデルである．DOM の木構造を用いて距離尺度を定義することも可能であるが，レイアウトのために深い入れ子構造を持つことが多く，表示されたときの近さとは必ずしも一致しないという問題がある．HTML をレンダリングした結果の座標情報などを用いれば，より正確な距離を定義できるが，レンダリングには計算コストがかかり，大量の Web ページを扱うには不向きである．そこで，本研究では DOM に含まれる HTML のブロックレベル要素について，直接テキストを含むのみを直列化（DOM の平坦化）して，その要素列内での位置関係により距離を定義する．

主要コンテンツの判定はテキスト量に基づく判定法 [2] を用いた．この手法を用いると主要コンテンツはかならず平坦化された要素列の連続する 1 区間（メイン区間）に対応する．このメイン区間との関係で以下の 2 種類の距離を定義する．

（1） メイン区間との距離（候補がメイン区間内にある場合は 0 と定義する）
（2） メイン区間を定義する境界との距離

3.3 著者らしさに基づくランキング

抽出された著者名候補には，前節で述べた主要コンテンツとの距離の他，著者名の言語的特徴（人名，組織名，未知語を含むなど），著者名の周辺のテキストや HTML タグなどが素性として与えられる．これらの素性を用いてランキングモデルを構築する．ランキングモデルには Ranking SVM [3] を用いた．

4．評価

データセットして情報信頼性評価用データ[1]に含まれる 2000 の Web ページのうち，サイト運営者とは異なる著者が 1 つだけあるような Web ページ 500 ページを用いた．これらのページには全て作業者により情報発信構成が付与されており，この中から著者名を抽出して正解データとした．評価方法として，出力されたランキングの 1 位が正解の場合，3 位以内に正解が含まれている場合，5 位以内が正解に含まれている場合，のそれぞれを正解とみなしたときの精度を用いた．

表 1: ランキングの精度

正解が含まれる順位	1 位	3 位以内	5 位以内
精度	48.6%	65.8	71.7

5．むすび

本稿では Web ページの著者の同定問題について，まず Web ページの情報発信構成の同定の部分問題であることを位置づけた．次に，Web ページから言語的特徴に基づき抽出された著者の候補の中から，文書構造中での主要コンテンツからの距離などの素性に基づく著者らしさのランキングモデルに基づき，著者を抽出する手法を提案した．実験の結果，上位 5 位以内で評価した場合に 70%を超える精度で抽出できることを示した．

提案手法では著者であることを判別するのに主要コンテンツからの距離を用いていることから，(1) 著者の抽出精度が主要コンテンツの判定精度に影響される，(2) ブログのプロフィール欄や著者紹介欄などが本文とは独立していて，文書構造上で著者名と主要コンテンツの距離が必ずしも近くない場合には精度が低い，といった問題がある．後者については著者紹介欄などを認識するといった方法などで解決できると考えられる．

情報発信構成全体の同定手法の実現に向けて，個別の発信者同定精度の向上と，発信タイプの判定手法の開発が今後の課題である．

参考文献

[1] Miyamori, H., Akamine, S., Kato, Y., Kaneiwa, K., Sumi, K., Inui, K., Kurohashi, S.: Evaluation data and prototype system wisdom for information credibility analysis. *Internet Research* 18(2):155–164, 2008.
[2] 加藤 義清，乾 健太郎，黒橋 禎夫 : Web ページの情報発信者の同定とその関係の抽出．言語処理学会第 14 回年次大会，pp.737-740，2008 年.
[3] Joachims, T.: Optimizing search engines using clickthrough data. In *Proceedings of KDD'02*, pp.133–142, 2002.