Webページの情報発信者の同定とその関係の抽出

Extracting the Identity and Relation of Information Senders from Web Pages

加藤 義清[*1], 乾 健太郎[*1], 黒橋 禎夫[*1*2],
Yoshikiyo Kato, Kentaro Inui and Sadao Kurohashi

[*1]情報通信研究機構
[*2]京都大学

1　はじめに

ブログや，Wiki，動画共有サイトなどが広く普及し，一般ユーザによる情報発信が盛んになる昨今，Web から得られる情報を活用するためには，その信頼性を適切に判断していくことがますます重要となっている．情報の信頼性を判断する上では，「誰が」「何を」言っているのかを正確に捉えることが必要となる．「何を」については，近年ブログなどでクチコミが広く書かれるようになった状況もあり，評判情報を抽出・解析する研究が盛んである ([5, 6] など)．「誰が」についてはオンライン書店におけるレビュアの評価機構などように，閉じたシステムの中で信頼できる情報発信者を評価していく取り組みがある一方で，一般の Web を対象に「誰が」を扱うものとして専門家検索 [1] や，Web から人間関係を抽出するソーシャルネットワークマイニング [3] などが挙げられるものの，情報発信者の分析という観点での研究はまだ少ない．

そこで，本稿ではまず Web ページの情報発信者を識別する問題に着目し，Web ページの情報発信者および情報発信者間の関係を情報発信構成として同定する問題として定式化する．情報発信構成は個々の Web ページについてそこから情報が発信されるにあたって，誰がどのような役割を持って関わっているかを捉えようとするものである．

次に，情報発信構成同定の実現に向けて，Web ページから情報発信構成の中心をなすサイト運営者を抽出する手法について述べる．提案手法では，情報発信者に関する情報がページ中の先頭や末尾に存在する傾向に着目し，HTML の構造を利用してページの先頭及び末尾に存在するテキストをまず抽出する．次に，Web ページから抽出されたサイト運営者名の候補をランキングする問題として捉え，出現頻度や品詞などを素性としたランキングモデルを機械学習により構築し，発信者名候補に適用する．ランキングモデルとして Ranking SVM[2] 法を利用した場合，トップにランクされたものを選択した場合の精度が約 45%，上位 5 番目以内に正解が含まれていれば正解とみなす方法で評価をすると精度が 60%以上と，単純に出現頻度が最も多いものを選択する方式に比べて，より高い精度でサイト運営者を抽出できる結果が得られた．

2　Web ページの情報発信構成

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人物や団体などを含む実体のことを意味する．そこには当然 Web ページの著者は含まれるが，それ以外にもその Web ページを公開している Web サイト[*1]の運営者や，Web ページの中で引用された情報の著者なども含む．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．

現在の検索エンジンでは直接得られない情報であるが，リテラシのあるユーザならば情報の信頼性を確認するときにその発信者やその背後にある組織を調べると言うことは当然のようにおこなっている．例えば，検索エンジンで検索結果と共に自動的に分析された情報発信構成を提供するだけでも，ユーザにとって十分に有益であると考えられる．更には，評価情報分析などの結果などと組み合わせることにより，発信者による意見の偏りを見るなど，信頼性を判断する上で有効だと考えられる，より高度な分析も可能となる．

2.1　情報発信構成の表現

情報発信構成は情報発信の形態を表す**情報発信タイプ**と情報発信者を記述する**情報発信者項**から構成される．例えば，「いろは産業」という企業のサイトにおいて，社長である「山田」氏の挨拶のページがあるとき，そのページの情報発信構成は次のように表される．

(所属，
　(営利団体:企業，株式会社いろは産業)，
　(-，山田太郎，社長，-))

ここで，1 項目の「所属」は情報発信タイプが「所属者発信タイプ」であることを示している．また，2 項目以降は情報発信者項である．このうち，最初の情報発信項が記述する発信者は，そのページにおけるサイト運営者であることを意味する．情報発信者が組織である場合，情報発信項は発信者クラスと名前で記述される．発信者クラスは図 1 で与えられたものを使用する．

*1　Web サイトをどのように定義するかは議論の余地があるが，ここではある特定の実体が編集権を有している Web ページ群で，かつそれらが一体のものとして認識されうる形で公開されている（同一のホストあるいは同一のディレクトリで提供されているなど）ものを Web サイトと考える．

1. 個人
   (a) 専門家
   (b) 準専門家
   (c) 一般
   (d) 匿名・ハンドル名
2. 団体
   (a) 営利団体
      i. 企業
      ii. 業界団体
2. 団体(続き)
   (b) 非営利団体
      i. 政府
      ii. 公益法人等
      iii. 政治団体
      iv. 大学
      v. 学会
      vi. 任意団体
   (c) 報道機関
      i. 新聞社
      ii. 出版社
      iii. テレビ・ラジオ
3. 不明

図 1: 情報発信者のクラス分類.

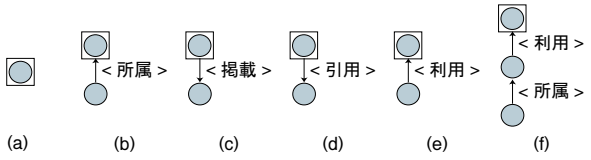

図 2: 情報発信構成で用いる 6 つの情報発信タイプ．丸いノードは情報発信者，エッジは情報発信者間の関係，四角に囲まれたノードはサイト運営者を表す．(a) 単一発信タイプ，(b) 所属発信者タイプ，(c) 掲載タイプ，(d) 引用タイプ，(e) サービスタイプ，(f) 複合タイプ．

(<発信者クラス>，<名前>)

個人の場合，発信者クラスと名前に加えて，職業・肩書と所属を記述する．ただし，所属は組織の情報発信項として表す．

(<発信者クラス>，<名前>，<職業・肩書>，<所属>)

2.2 情報発信タイプ

情報発信構成では図 2 に示す 6 種類の情報発信タイプを用いる．以下，それぞれについて説明する．

単一発信者タイプ　Web ページの情報発信者が単一である場合，そのページの情報発信構成は**単一発信者タイプ**とする．単一発信者タイプの場合，サイト運営者がそのページにおける唯一の情報発信者となる．例えば,「山田花子」という人の個人的なホームページのあるページを対象とした場合，情報発信構成は次のようになる．

(単一，(個人:一般，山田花子))

所属発信者タイプ　Web ページの情報発信者が複数いる場合で，全ての発信者がサイト運営者自身であるか，サイト運営者に所属する人物である場合，そのページの情報発信構成は**所属発信者タイプ**とする．所属関係は他の関係（掲載や引用）に優先する．例えば，ある企業の社長が，雑誌のインタビュー記事で取り上げられて，その内容がその企業のサイトで掲載されている場合，その企業と社長には掲載関係があるが，所属関係が優先するので，情報発信構成は所属発信者タイプとなる．所属発信者タイプの場合，所属者の情報発信項の中で，発信者クラスや所属は与えない．所属発信者タイプの例については 2.1 節の例を参照のこと．

掲載タイプ　Web ページでサイト運営者とは異なる第三者の著作が掲載されている場合，そのページの情報発信構成を**掲載タイプ**とする．掲載タイプは情報の内容について，主に被掲載者に文責がある場合を想定している．例えば，学会のサイトに掲載された論文や，新聞社のサイトに掲載された専門家の寄稿記事などである．掲載タイプの場合，非掲載情報の著者は，そのページに自らの著作物が掲載されることを承知している場合を想定する．次に述べる引用タイプとの違いに留意されたい．

(掲載
  (報道機関:新聞社，○×新聞社)，
  (個人:専門家，鈴木一郎，大学教授，
    (非営利団体:大学，海山大学)))

引用タイプ　Web ページの中で，第三者による情報を引用している場合，そのページの情報発信構成を**引用タイプ**とする．被引用情報以外に引用者で発信する情報がページの主な情報となる場合は引用タイプとなる．被引用情報は引用者の責任で引用されており，被引用者が引用されていることを必ずしも承知していない場合を想定している．引用タイプと掲載タイプの違いは Web ページの目的と，主たる情報発信者（掲載や引用をする側）の意図による．第三者の著作物を掲載し，それを世の中に広めることを目的とする場合には掲載タイプである一方で，引用者に何らかの主張があって，それを補強するために第三者の情報を利用している場合には引用タイプとなる．例えば，あるブログで新聞記事を引用している場合などは，引用タイプとなる．

(引用，
  (個人:一般，山本さくら，-，-)，
  (報道機関:新聞社，○×新聞社))

サービスタイプ　掲示板やブログのコメント欄などように，サイト運営者によってサイト運営者以外のものが比較的自由に情報を発信できるように作られている Web ページの場合，その情報発信構成は**サービスタイプ**とする．例えば,「一二三株式会社」が運営する掲示板サイトで,「じゅん」というハンドルネームで投稿がある場合は次のように記述する．

(サービス，
  (営利団体:企業，一二三株式会社)，
  (個人:匿名，じゅん，-，-))

複合タイプ　一つのページの中でこれまでに述べた情報発信タイプが複合して現れる場合に情報発信構成を**複合タイプ**として記述する．例えば，ある小売店のサイトで，店長が購入者のコメントを紹介している場合には，小売店と店長との間には所属関係があり，かつ店長と購入者の間には引用関係があると考え，次のように記述する．

(複合，
  (所属，
    (営利団体:企業，○×ダイエットショップ)，
    (-，柏原，店長，--))，
  (引用，

(個人:一般，柏原，店長，
(営利団体:企業，○×ダイエットショップ))，
(個人:一般，小野，-，-)))

3 サイト運営者の抽出

情報発信構成の同定に向けて，まずサイト運営者を抽出することが重要となる．それは，サイト運営者が，サイト運営者以外の発信者との関係を知る上で重要な手掛かりとなるからである．例えばサイト運営者が新聞社であると分かれば，情報発信構成は掲載タイプとなる傾向にあることが予想される．そこで，本稿では情報発信構成の核となるサイト運営者の抽出手法について述べる．サイト運営者の抽出は以下の手順でおこなわれる．

(1) 情報の収集
(2) HTML からの発信者名抽出領域の選択
(3) 選択領域からの候補テキストの抽出
(4) 候補テキストからサイト運営者名候補の抽出
(5) サイト運営者名のランキング

3.1 情報の収集

情報発信者に関する情報は分析対象のページ内のみに存在するのではない．例えば，多くのサイトでは「会社概要」や「プロフィール」といった情報発信者に関する情報が掲載されているページが設けられており，他のページからリンクが張られている．また，サイト運営者が Web サイトのドメイン名を管理しているような場合には，WHOIS データベースを検索することによってサイト運営者に関する情報が得られる．

ここでは，分析対象ページ以外に，1) ルートページ (toppage) を含む祖先ページ (ancestor)，2) 特定の文字列（「会社概要」など）を含むアンカーテキストでリンクされているページ，3) WHOIS データベースを情報源として利用する．以降，それぞれの情報源を一つの文書とする文書集合からサイト運営者名を抽出する．

3.2 抽出領域の選択

抽出対象となったページに対して，図 3 に示したアルゴリズムを適用し，ページの先頭および末尾の領域を抽出対象領域として選択する．基本的な考え方は，発信者に関する情報はページの先頭および末尾近くに記載されていることが多いという性質を利用するというものである．まず，HTML を DOM 木に変換し，body 要素を起点として以下の処理を再帰的に適用する．子ノードについてそれぞれをルートとする部分木に含まれるテキストの量のページ全体のテキスト量に対する割合を算出し，ある閾値以上ならばページの本体部分 (図 4 の Main Block) とみなす (図 3 の MainBlock)．本体部分とされた以外のノードが含むテキストは全て抽出対象となる．本体部分とされたノードに対しては MainBlock を再帰的に適用する．MainBlock となる子ノードを持たないノードに至った時点で，そのノードのテキスト量全体に対する割合に基づいて，先頭と末尾に該当するノードを抽出対象として選択する (図 4 の Upper と Lower)

body 要素に含まれるテキスト以外に，head 要素にある title 要素のテキスト, meta 要素のうち name 属性

Algorithm 3.1: ExtractText(DOM)

procedure MainBlock(n)
$Ext = \phi$
$l_n =$ TextLength(n)
$C \leftarrow$ Children(n)
$main \leftarrow \phi$
for each $c_i \in C$
do { $l_i \leftarrow$ TextLength(c_i); **if** $l_i/l_n > t_m$ **then** { $main \leftarrow c_i$; *exit loop* }
if *main exists*
then { **for each** $c_i \in C/\{main\}$ **do** $Ext \leftarrow Ext \cup$ Text(c_i); $Ext \leftarrow Ext \cup$ MainBlock($main$) }
else $Ext \leftarrow Ext \cup$ HeaderFooter(n)
return (Ext)

main
$body =$ Element(DOM, body)
return (MainBlock($body$))

図 3: 発信者名抽出領域アルゴリズム.

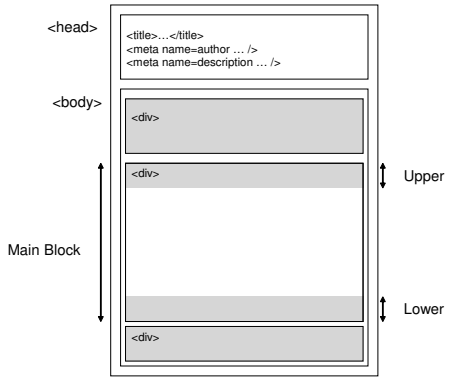

図 4: 発信者名抽出領域の選択．網掛された部分が抽出対象領域.

が author, description, keywords であるもの content 属性も候補テキストとして抽出する．

3.3 候補テキストの抽出

抽出対象領域に含まれるテキストに対して，文に分割した上で次の条件を満たすものを候補テキストとして残す．(1) 文に含まれる助詞のうち「の」以外の助詞の割合がある閾値以下である (2) 人名，組織名，地名，組織名末尾，未定義語のいずれかが含まれる

3.4 情報発信者候補の抽出

候補テキストから複合名詞を抽出して，情報発信者名の候補とする．複合名抽出は，入力テキストに KNP による構文解析を適用し，解析結果からルール[*2]に基づいておこなう．

3.5 サイト運営者のランキング

ここまでに述べた方法で抽出された情報発信者候補の中から，次の素性に基づくランキングのモデルを訓

*2 例えば,『文節内で連続する「名詞相当語」とラベル付けされた形態素列』など.

練データより機械学習により構築する．用いる素性は，1) 情報源全体における出現頻度 (tf)，2) 候補が出現する文書の頻度 (df)，3) 候補が出現する文書の種類（分析対象ページ (target)，トップページ (toppage)，祖先ページ (ancestor),「会社概要」「プロフィール」など特定のアンカーテキストでリンクされたページ (about))，4) 構成語の品詞属性（「組織名」,「組織名末尾」「人名」「地名」），5) 先頭形態素・末尾形態素，6) 形態素数，である．ランキングのモデルには，Ranking SVM[2] を用いる．

3.6 評価

評価用のデータとして情報信頼性分析評価用データ [4] の 20 トピック（約 2000 ページ）の内，18 トピック分 (1751 ページ) を利用した．これらの各ページに対して 1 人の作業者が与えた情報発信構成に基づいて，訓練および評価事例を作成した．具体的には，3.4 節で抽出された発信者名候補について，情報発信構成から取り出されたサイト運営者の名前と比較して，名前が完全に含まれる事例には 2，ある基準（元の名前に対する一致した部分文字列の長さの割合など）を満たした上で部分的に含まれる事例に対しては 1，それ以外の事例については 0 をそれぞれ Ranking SVM [*3]のラベルとして与え，同じページから抽出された候補の間で優先度の制約を与えるようにした．ベースラインとして，抽出された発信者名候補のうち最大の tf を与えるものをサイト運営者とした場合の精度を評価したところ，32.7%であった．

評価は各ページについて，ランキングされた候補の中で上位 k 番目以内に正解発信者名が含まれていればそのページは正解としたときの精度でおこなった．先に述べたデータを用いて，18 分割交差検定（トピック毎に分割）を実施した結果を図 5 に示す．All は発信者名候補のいずれかに正解が含まれている場合に正解とした場合である．1.0 となっていないのは，抽出された候補の中に正解が含まれていないケースがあるためである．すなわち，All の場合の値が抽出精度の上限を示すことになる．いずれのケースでも提案手法がベースラインを上回っているが，特にランク順位が 1 番目のものだけで評価するケースで性能の改善が大きい．

提案手法で誤りが多かったケースとして，(1) ブログにおいて発信者の名前がハンドル名のみの場合，(2) トピックが特定の組織についてのものだった場合，が挙げられる．(1) のケースではハンドル名には人名や組織名といった手掛かりになる品詞属性がつかないものも多く，そのために候補の中に含まれていても上位にランクされないことが多かった．(2) のケースでは，トピックに組織名が含まれる場合，出現頻度も多くなるので，組織名であることとの相乗効果により，誤ってトップにランキングされることが多かった．これら問題は，現在のモデルが候補名の言語的な特徴に偏った形で学習されているためだと考えられる．今後，発信者名候補の出現位置や，出現のコンテキスト（例えばブログの場合,「Posted by〜」などのように，発信者を端的に表す記述があることが多い）など，tf や df 以外の非言語的特徴も利用することを検討する．

*3 Ranking SVM 法の適用には SVM Light (http://svmlight.joachims.org/) を用いた．

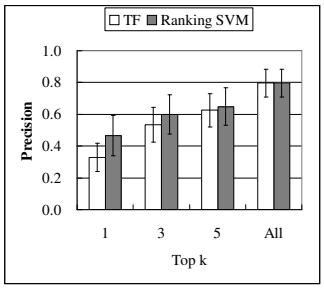

図 5: 上位 k 番目までに正解が含まれる場合を正解としたときの精度

4 おわりに

本稿では，Web ページの情報発信者の分析について，情報発信者構成を同定する問題として定式化し，情報発信者構成の中心となるサイト運営者を抽出するための手法を提案した．評価の結果，ベースラインとして設定した抽出された発信者名を頻度順で並べた場合に比べて，トップにランキングされたものだけで評価した場合に 1.4 倍以上の精度向上が確認できた．情報発信構成の同定に向けて，情報発信者の情報発信クラスおよび情報発信タイプの推定が今後の課題である．

謝辞

本研究を進めるにあたってご協力頂きました情報通信研究機構の赤峯享，河原大輔，森井律子の諸氏，ならびに京都大学の柴田知秀氏に感謝の意を表します．

参考文献

[1] I. Becerra-Fernandez. Searching for experts on the web: A review of contemporary expertise locator system. *ACM Transactions on Internet Technology*, 6(4):333–355, 2006.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, New York, NY, USA, 2002. ACM.

[3] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hashida, and M. Ishizuka. Polyphonet: An advanced social network extraction system from the web. In *Proceedings of WWW2006*, pp. 397–406, 2006.

[4] H. Miyamori, S. Akamine, Y. Kato, K. Kaneiwa, K. Sumi, K. Inui, and S. Kurohashi. Evaluation data and prototype system wisdom for information credibility analysis. In *Proceedings of the First Workshop on Information Credibility on the Web*, pp. 25–32, 2007.

[5] N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin. Computational approaches to analyzing weblogs: Papers from the 2006 spring symposium. Technical Report SS-06-03, American Association for Artificial Intelligence, Menlo Park, California, 2006.

[6] 奥村. blog マイニング：インターネット上のトレンド，意見分析を目指して. 人工知能学会誌, 21(4):424–429, 2006.