Web 文書の情報発信者クラス分類

Classifying Information Sender of Web Documents

加藤 義清[*1], 黒橋 禎夫[*2*1], 乾 健太郎[*3*1]
Yoshikiyo Kato, Sadao Kurohashi and Kentaro Inui

[*1]情報通信研究機構
[*2]京都大学
[*3]奈良先端科学技術大学院大学

1 序論

World Wide Web の登場により，一般の人が世界に対して情報発信することが容易となった．特に近年，ブログや SNS (Social Network Service) の普及により，情報発信のコストが下がり，口コミ情報など User Generated Content (UGC) や Consumer Generated Media (CGM) などと呼ばれる，今までは得られなかった種類の情報が大量に取得可能な状況が現れている．このように発信される情報の多様性が増す中で，PageRank[2] など，リンクを一種の人気投票と見なしてページの重要度を計算する一次近似的な手法はあるものの，Web から得られる情報の信頼性を一般の利用者が判断するための支援は限られている．

本研究では情報の信頼性は発信者，皮相，意味，評判という 4 つの観点により評価できるという立場を取る [3]．本稿はこの 4 つの観点の内，情報発信者の諸特性に基づく情報の信頼性評価に係わる手法について述べる．

ある情報の信頼性を判断する際，その情報が誰により発信されたのかというのは非常に重要な要素の一つとなる．Web 文書に記述される情報の信頼性についても同様である．しかし，現実世界に比べ Web の世界では誰が発信した情報なのかということが非常に分かりにくくなっている．

本稿では，情報発信者に基づく信頼性評価への第一歩として，Web サイトの発信者クラス分類の問題について，サイトのトップページの情報に基づく簡便な分類手法とその評価について報告する．評価の結果，本稿で報告する手法により，80%の精度で Web 文書の情報発信者のクラスを推定することができることが明らかとなった．ただし，評価に用いたデータについて，分類が比較的簡単な発信者クラスが多数を占めており，そのクラスを除いた分類精度は 40%を切った．今後の改良に向けて，分類がうまくいかなかった場合についての考察について述べる．

2 分類体系

本研究では，発信される情報の信頼性を判断することを目的として情報発信者の分類をおこなおう．従って，分類の結果として得られる情報発信者のクラスは，その判断の助けとなるものでなくてはならない．本研究ではこの考え方に基づいて，表 1 に掲げる分類体系を用いた．

表 1: 情報発信者の分類体系

1. 個人 (I)
   - (a) 有識者・専門家・著名人
   - (b) 一般
   - (c) 匿名・ハンドルネームのみ
2. 団体
   - (a) 営利団体 (PO)
     - i. 企業
     - ii. 業界団体
   - (b) 非営利団体
     - i. 行政 (AD)
     - ii. 公益法人・NPO 等 (NPO)
     - iii. 大学 (U)
     - iv. 学会 (AC)
     - v. 任意団体 (VOL)
   - (c) 報道機関
     - i. 新聞社 (NP)
     - ii. 雑誌 (MG)
     - iii. テレビ・ラジオ等 (BC)

まず，個人であるか団体であるかの区別をおこなう．個人の中では有識者，専門家など何らかの専門知識を有すると考えられる人と，一般の人，および匿名の場合とを区別した．団体については，その情報発信の意図を考えるときに，団体が営利を追求しているかどうかは重要な要素であると考え，まず営利団体と非営利団体に分けている．同じレベルに報道機関をおいているのは，報道機関も基本的には営利企業であって営利団体に分類されるるが，それらの発信する情報の信頼性は特別に考慮されるべきものだとして，別に項目を立てた．

非営利団体について，大学や学会といった項目は (2-b-ii) 公益法人・NPO 等に含めても良いが，個人で有識者や専門家を区別したのと同様，専門知識を有している情報発信主体であるとして，別に項目を立てた．

3 手法

次に，Web サイトのトップページに基づく情報発信者分類手法について述べる．本手法は，Web サイトのトップページの URL と HTML ファイルの<title>タグの情報（タイトル），及びトップページへのハイパーリンクに含まれるアンカーテキストを用いて情報発信者クラスを推定する．本手法により与えられる分類は，表 1 の右に括弧付きでアルファベットが記載されている項目（I, PO 等) である．本手法は，1) 接辞処理などの前処理，2) 手がかりの抽出，3) 手がかりを用いた分類，という 3 段階を経て発信者クラスを与える．

以下，それぞれの段階について述べる．

3.1 前処理

本手法では前処理として，分類に用いるタイトル及びアンカーテキストの接辞処理をおこなう．これは，「〜のホームページ」のように，タイトルやアンカーテキストにおいて，情報発信者の名前を含む頻出パターンを想定したものである．このように発信者名のまわりに現れる「のホームページ」などの接辞を前処理に取り除くことにより，発信者名を抽出し，発信者クラスの分類の精度が向上することを意図している．

以下に述べる処理は全て，処理対象文字列を日本語形態素解析システム JUMAN[4][*1]により形態素列化したものに対しておこなう．

入力形態素列 $I = \{m_i\}$, 接頭辞パターン集合 $\mathcal{P} = \{P_i\}, P_i = \{p_{ij}\}$，接尾辞パターン集合 $\mathcal{S} = \{S_i\}, S_i = \{s_{ij}\}$ が与えられたとする．ただし，m_i, p_{ij}, s_{ij} は形態素である．

1. 接頭辞の除去
   $I = \{m_i\}$ について，$m_i = p_{ji}$ $(p_{ji} \in P_j,\ i = 1..|P_j|)$ となるような $P_j \in \mathcal{P}$ のうち，その長さ $|P_j|$ が最大のものについて，I の先頭から除去する．
2. 接尾辞の除去
   $I = \{m_i\}$ について，$m_i = s_{ji}$ $(s_{ji} \in S_j,\ i = |I| - |S_j| + 1..|I|)$ となるような $S_j \in \mathcal{S}$ のうち，その長さ $|S_j|$ が最大のものについて，I の末尾から除去する．
3. I の長さが変わらなくなるまで 1-2 を繰り返す．

接辞パターン集合については，評価でも用いた NTCIR-5 Web タスクテストコレクション [1] の 50 万ページ分のタイトルおよびアンカーテキストについて，n=1〜5 について，接頭辞，接尾辞それぞれについてパターンを数え上げ，頻出パターンのうち，発信者の情報を含まない除去すべきパターンのみを残すことによって得た．表 2 に接辞パターンの例を示す．

3.2 手がかり抽出

本手法で用いる手がかりの概要について表 3 に示す．大きく分けて，ページの URL から得られるものと，タイトルやアンカーテキストなどの文字列から得られるものがある．URL からは Domain(C) の 4 種類，文字列からは Individual, LastMorpheme および MorphemePattern(C) を合わせて 10 種類の手がかりが得られる．アンカーテキストについては，一つのページが複数持つことがあるので，各手がかりの合計値を用いる．ただし，LastMorpheme については手がかり値がクラスとなるので，クラスを数え上げてそれぞれの個数を新たな手がかりとする．このようにして，全体では 34 個の手がかりを抽出する．

*1 以降,「形態素解析をおこなう」といったときには JUMAN を利用して形態素解析をおこなうことを意味する．

表 2: 接辞パターンの例

接頭辞	接尾辞
Welcome	の:ホームページ
ようこそ	公式:ホームページ
トップ	に:ようこそ
ホームページ	Web: :Site
トップ:ページ	の:ホームページ:へ:ようこそ:！
Welcome: :to	Official: :Web: :Site

(注)':' は形態素の区切りを示す．

3.3 分類

抽出された手がかりに基づいて，以下の順で何らかのクラスが与えられるまで分類規則を適用する．1-6 までにクラスが与えられない場合には，未分類 (UC) とする．

1. Domain(C)=1 であるようなクラス C を分類として与える．例えば，Domain(PO)=1 ならば，PO と分類する．
2. タイトルの LastMorpheme で与えられるクラスを与える．
3. アンカーテキストの LastMorpheme で与えられたクラスのうち，最も多かったクラスを与える．
4. タイトルに対する MorphemePattern(C) で値が最も大きかったクラスが分類として与えられる．
5. アンカーテキストに対する MorphemePattern(C) で値が最も大きかったクラスが分類として与えられる．
6. Individual=1 ならば I をクラスとして与える．
7. クラスとして未分類 (UC) を与える．

4 評価

4.1 方法

NTCIR-5 Web タスクテストコレクション [1] の中から任意の 50 万ページについて，その中に含まれるトップページについて前節で述べた発信者分類手法を適用し，発信者クラスの分布，被覆率，および分類精度を調べた．トップページかどうかの判断は，ページの URL のパスがルート ('/') になっている場合にトップページであると判断した．その結果，50 万ページ中，2492 ページがトップページと判断され，評価の対象となった．

4.2 結果

図 1 は分類の判断に利用した URL，タイトル，アンカーテキストについて，手がかりが与えられるものを数え挙げて，全体に占める割合を示したものである．この結果から，全体の 8 割のサイトは URL の手がかりにより何らかの分類が可能であることが分かる．一方，URL に比して，タイトルおよびアンカーテキストに手がかりが含まれる割合は低いが，3 つの情報全てを組み合わせることにより，URL のみを用いる場合と比べて，より多くのサイトで手がかりが得られる，約 90%のサイトに対して分類を与えることができることが分かった．

表 3: 発信者分類に用いる手がかり

手がかり	対象クラス	判定基準
Domain(C)		ドメイン名が対応付けられたパターンのいずれかにマッチする場合，1 とする．
	PO	.co.jp, .com
	NPO	.or.jp, .org, .ed.jp
	AD	.go.jp, .gov
	U	.ac.jp, .edu
Individual	I	入力文字列に対して形態素解析をおこない，品詞「名詞:人名」に分類される形態素を含んでいれば 1 を与える．
LastMorpheme		入力文字列に対して形態素解析をおこない，末尾の形態素が意味素に「組織名末尾」を含む場合，形態素に対応付けられたクラスを手がかり値とする．「組織名末尾」を意味素に含む形態素については全て予めクラスとの対応付けがなされている．
	VOL	「委」「院」「会」「会議」など
	PO	「銀行」「興業」「工業」「鉱業」「航空」「支社」「支店」など
	NPO	「学院」「学園」「協会」「高校」「高専」「公団」「小学校」など
	AD	「艦隊」「機構」「基地」「局」「議会」「軍」「警察」「県警」など
	U	「医大」「大」「大学」
	AC	「学会」
	BC	「テレビ」
MorphemePattern(C)		入力文字列に対して形態素解析をおこなって得られる形態素列が，特定の文字列集合に形態素解析を施すことにより得られる形態素列集合に含まれる形態素列を含む個数を手がかり値として与える．
	PO	「(株)」「株式会社」「サービス」「製造」「商事」など 66 個
	NPO	「公益法人」,「財団」,「(社)」,「医療法人」,「学校法人」など 29 個
	AD	「国立」「県立」「府立」「道立」「都立」「独立行政法人」
	U	「大学」「研究室」「Lab」「lab」「laboratory」「Laboratory」
	AC	「学会」「Society」「Association」「Institute」
	NP	「新聞」
	MG	「週刊」「月刊」
	BC	「ラジオ」「テレビ」

次に，表 4 に，分類手法を適用した結果，各クラスに分類されたサイトの分布を示す．この表により，今回用いたデータセットの発信者の傾向が分かる．8 割以上を営利団体が占めており，サイト単位で見れば発信者の分布に大きな偏りがあると言える．

表 5 に，サンプリングによる各手法の被覆率と精度を評価した結果である．全データ 2492 件から 100 件分をランダムサンプリングして，人手で分類を与えた．このうち，人手で分類が与えられたサイト (有効サンプル) について，人が与えた分類と，各手法での分類結果を比較し精度を算出した．被覆率は，有効サンプルの内，各手法により分類が与えられたサイトの割合を算出したものである．この結果では，約 80%の精度で分類できる結果となった．

表 6 に全データのうち，PO 以外に分類されたものについて精度を評価した結果を示す．このような評価をおこなった理由は，以下のようなものである．実験に用いたデータセットの 8 割近くを PO が占めるが，PO については URL に基づく分類により比較的高い精度で分類できたと予想される．一方，表 5 は全体からのサンプリングであるため，PO の割合が多く，PO 以外の分類精度を十分に表さないと考えためである．この評価により，全体の場合に比べ，PO 以外の場合では精度が全体で評価した場合の精度の 1/2 程度に落ち込むことが明らかとなった．

4.3 誤り分析

誤りの例についていくつか見てみる．

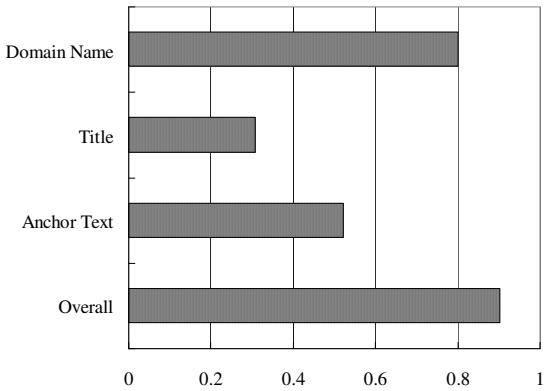

図 1: 各手がかりを有するトップページの割合．Overall はいずれかの手がかりを有しているトップページの数

誤って I と分類される場合

クラス I は手がかり Individual に基づく分類規則より与えられるが，他の分類規則に比べて優先順位が低いので，その他の手がかりは与えられなかったことになる．そのような中で良く見られた誤りは，(1) 地名が人名とし誤認識される場合，及び (2) 組織名に人名を表す語が含まれている場合である．(1) について，JUMAN で「山梨」「浦和」「渋谷」という語に付与される品詞として第一候補は人名となっている．そのため，タイ

表 4: 分類された発信者クラスの分布

Class	Count	Percentage
I	53	2.1
VOL	6	0.2
PO	2096	84.1
NPO	74	3.0
AD	5	0.2
U	8	0.3
AC	0	0.0
NP	2	0.1
MG	0	0.0
BC	7	0.3
UC*1	241	9.7
Total	2492	100.0

*1: 未分類サイト

表 5: 分類の被覆率と精度

項目	値
有効サンプル数 (N_e)	90
分類サイト数 (N_c)	78
正解数 (N_p)	63
被覆率 (N_c/N_e)	0.87
精度 (N_p/N_c)	0.81

トルやアンカーテキストに地名が含まれていた場合に誤って I と分類される例が見られた．(2) については，「オートショップ中村」のように店の名前などに人名が含まれる場合である．

誤って NPO と分類される場合

NPO は Domain, LastMorpheme, MorphemePattern のいずれの手がかりによっても分類される可能性がある．誤って NPO と分類された場合でよく見られた例は，(1) .org および.or.jp のドメインを有するサイトが NPO 以外に分類されるべきサイトである場合，(2) LastMorpheme や MorphemePattern に使われるパターンについて NPO でないのに現れる場合である．(1) について，.org や.or.jp が I や PO など，NPO 以外のサイトのドメインで使われる例が見られ，発信者クラスを推定するのに必ずしも有効でないことが分かった．(2) について,「協会」という語について LastMorpheme と MorphemePattern でともに NPO と関連づけられているが,「協会」が文字列に含まれていても，NPO に分類されないサイトがあった．例えば，業界団体や任意団体などが「〜協会」と名乗っている例が見られた．NPO に関しては分類に用いられるパターンが正確でないために誤っていると言える．

4.4 考察

今回の結果より，URL，タイトル，アンカーテキストの解析による簡便な方法により，比較的高い精度で発信者クラスを分類することが可能であることが明らかになった．但し，比較的高い精度が得られたのは，PO の場合，'.co.jp' がドメイン名に含まれることにより容易に分類でき，更には PO と分類されるサイトがデータ中に占める割合が高かったためだと考えられる．一方，表 6 でも明らかなように，PO 以外の分類では十分な精度が得られていない．誤り分析でも見たように，分類に用いられるパターンが十分に洗練されていないことから，今後パターンを修正していくことにより，ある程度改善が見込まれると考える．

表 6: PO 以外のクラスに分類されたサイトについての分類精度

項目	値
PO 以外に分類されたサイト数	155
分類可能サイト数*1 (N_e)	148
正解数 (N_p)	58
精度 (N_p/N_e)	0.39

*1: PO 以外に分類されたサイトのうち，人間により分類を与えることができなかったものを除いた数.

それ以外の要因として，図 1 からも分かるように，タイトルとアンカーテキストから手がかりを抽出できるのは全体の 30〜50%であり，PO の場合のように URL のみで高精度の分類ができない PO 以外のクラスでは，手がかりが不足していることにより精度が上がらないと予想される．今回用いた手法では，トップページの本文や，サイト内の他のページの情報を全く用いていないため，今後それらの情報も分類に利用することにより，被覆率および，特に PO 以外の分類精度の向上が期待される．

5 結論

本稿では，Web 文書の情報発信者を分類ための URL，タイトル，アンカーテキストを用いた手法と，その評価について報告した．全体の精度は約 80%であったが，URL を見ることにより容易に分類することができる PO がデータ中に占める割合の大きさが，大きく貢献しており，PO 以外の分類については約 40%の精度にとどまった．その原因の一つとして，タイトルとアンカーテキストのみでは発信者に関する情報量が不足していることが挙げられる．

今後の課題として，トップページの本文や，同じサイトのトップページ以外の情報も用いることが挙げられる．

また，本稿で述べた手法は一つのサイトに一つの発信者を想定して分類をおこなっており，blog サービスなどのように，同じドメインでも複数の発信者がいる場合には対応できていない．同一サイト内で，複数の発信者を識別し分類することも今後の課題である．

参考文献

[1] Kando, N. and Takaku, M.(eds.): *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (2005).

[2] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford University (1999). [Available at http://dbpubs.stanford.edu/pub/1999-66].

[3] 加藤義清, 黒橋禎夫, 江本　浩:情報コンテンツの信頼性とその評価技術, 人工知能学会研究会資料, SIG-SWO-A602-01 (2006).

[4] 黒橋禎夫, 河原大輔:日本語形態素解析システム JUMAN version 5.1，東京大学大学院情報理工学系研究科 (2005).