Web ページの情報発信構成の同定

加藤 義清[1)]，河原 大輔[1)]，乾 健太郎[1), 2)]，黒橋 禎夫[1), 3)]
[1)]情報通信研究機構　[2)]東北大学　[3)]京都大学
{ykato,dk}@nict.go.jp　kuro@i.kyoto-u.ac.jp

1　はじめに

近年，Web ページから評判情報を抽出する意見解析の研究 [4, 5, 3] や，Web ページによく現れる言説を抽出して提示する情報要約の研究 [1] など，Web ページに書かれた内容の解析に関する研究が盛んである．その結果を利用者に提示するとき，意見や要約の内容と同様に重要になってくるのが，誰がその内容を述べたのかという，情報の発信者についての情報である．

本研究では，Web 上の情報の信頼性を分析するという文脈において，Web ページの情報の発信者の問題を扱う．情報の信頼性を分析するといったときに様々なアプローチがあり得るが，著者らは分析の対象となる主題について，複数の情報源からの関連情報を収集した上で，様々な観点からの分析を施し，その結果を要約して俯瞰的な情報として提示することにより，利用者が信頼性を判断しやすくするというアプローチを採る [8]．本研究で扱う情報発信者は，この「様々な観点からの分析」の一つである．

情報発信者と一言言っても，Web ページの情報発信には様々な主体が様々な役割で関わっており，問題は見かけほど単純ではない．著者らは既に Web ページの情報発信者とその関係を情報発信構成として捉えることを提案している [6]．Web ページの情報発信構成は，(1) Web ページの情報発信者，(2) 情報発信者の分類を与える情報発信者クラス，および (3) 情報発信者の役割や情報発信者間の関係を表す情報発信タイプから構成される．Web ページの情報発信者については，[6] において同定手法を提案した．本稿では情報発信者に加えて，情報発信者クラスおよび情報発信タイプを同定することにより，情報発信構成全体を同定する手法について述べる．

本稿の構成は次の通りである．まず，次節において情報発信構成について説明する．3 節では，Web ページの情報発信構成を同定する方法として，情報発信者の同定と関係の同定を逐次的におこなう手法と，情報発信者の同定と関係の同定を同時におこなう手法の 2 つの手法を提案する．4 節において提案手法による実験について報告した後に本稿を締めくくる．

2　Web ページの情報発信構成

Web ページによる情報の発信においては，発信される情報の著者以外にも複数の主体（サイト運営者など）が，それぞれ異なる役割（出版，引用など）で情報発信に関わっていることが多い．そのため，複数の情報発信者がいるような場合に Web ページの情報発信者と言ったとき，どの情報発信者のことを指すのか必ずしも自明ではない．本研究では，Web ページの複数の情報発信者を，その役割や情報発信者間の関係も含めた上で認識し，記述するための方法として情報発信構成の考え方を導入している．本節では，情報発信構成について簡単に説明する．詳細については [6] を参照されたい．

情報発信構成は情報発信者，情報発信者クラス，情報発信タイプの 3 つの要素により構成される．

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人物や団体などを含む実体のことを意味する．情報発信構成においては，次の 2 種類の情報発信者を区別する．**サイト運営者**は Web ページを公開している Web サイトの運営者である．**著者**は Web ページ内で公開されている情報の著者である．Web ページの情報発信構成とは，サイト運営者と著者，各情報発信者の発信者クラス，および情報発信者間の関係を与えるものである．

情報発信タイプは情報発信者間の関係や役割を示すもので，本研究では図 1 に示す 6 種類の情報発信タイプを定義している．**情報発信者クラス**は，情報発信者が組織であるか個人であるか，営利目的か否かなど，いくつかの観点から分類したものであり，本研究では約 20 のクラスを定義している．

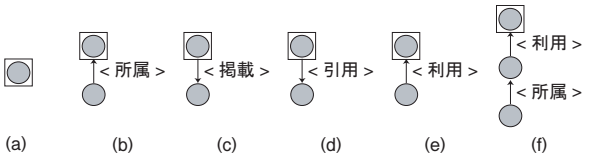

図 1: 情報発信構成で用いる 6 つの情報発信タイプ．丸いノードは情報発信者，エッジは情報発信者間の関係，四角に囲まれたノードはサイト運営者を表す．(a) 単一発信者タイプ，(b) 所属発信者タイプ，(c) 掲載タイプ，(d) 引用タイプ，(e) サービスタイプ，(f) 複合タイプ．

情報発信構成は，ここまでに述べた要素を用いて以下の形式で記述される．

(情報発信タイプ，サイト運営者 [，著者]*)

サイト運営者および著者は情報発信者項として記述される．情報発信者項は，情報発信者の情報発信者クラス，名前，肩書き，所属などを記述したものであり次のような形式を取る．

(情報発信者クラス，名前 [，肩書き，所属])

図 2 は情報通信研究機構のサイトのページで，理事長である宮原秀夫氏の挨拶を掲載したものである．このページの情報発信構成は以下のように記述される．

(所属，
(政府機関，独立行政法人情報通信研究機構)，
(-，宮原秀夫，理事長，-))

この記述から，このページには情報発信者が 2 つ存在し，サイト運営者が「独立行政法人情報通信研究機構」，著者が「宮原秀夫」であり，著者がサイト運営者に所属しているということが分かる．

3 情報発信構成の同定

情報発信構成の同定は Web ページが与えられたときに，情報発信構成の記述を求める問題となる．本稿では，Web ページの情報発信構成を同定する方法として，情報発信構成の各要素を逐次的に同定する手法（逐次同定法）と，全ての要素を同時に同定する手法（同時同定法）を提案する．なお，本稿では情報発信者の数がたかだか 2 であるような情報発信構成について扱うものとする．

図 2: Web ページの情報発信者

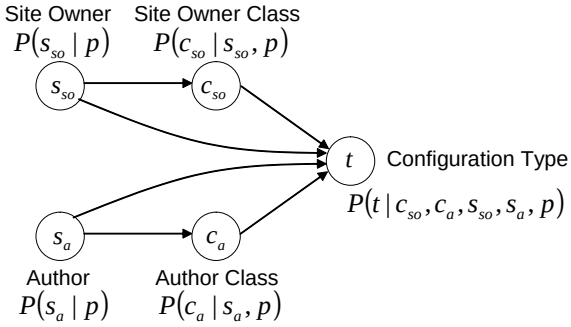

図 3: 情報発信構成の逐次同定モデル．

3.1 情報発信構成の逐次同定法

逐次同定法においては，サイト運営者 (s_{so})，著者 (s_a)，サイト運営者の情報発信者クラス (c_{so})，著者の情報発信者クラス (c_a)，および情報発信タイプ (t) に図 3 に示すような依存関係があると仮定し，1) s_{so} と s_a，2) c_{so} と c_a，3) t という 3 段階で各要素を同定していく．具体的には，前段階の同定結果を次段階の同定の素性として用いつつ，以下のタスクを順に実行することにより，情報発信構成 $< t, s_{so}, s_a, c_{so}, c_a >$ の各要素を逐次的に同定する．

$$s^*_{so} = \arg \max_{s_{so} \in S_{so}(p)} P(s_{so}|p) \quad (1)$$

$$s^*_a = \arg \max_{s_a \in S_a(p)} P(s_a|p) \quad (2)$$

$$c^*_{so} = \arg \max_{c \in \mathcal{C}} P(c|s^*_{so}, p) \quad (3)$$

$$c^*_a = \arg \max_{c \in \mathcal{C}} P(c|s^*_a, p) \quad (4)$$

$$t^* = \arg \max_{t' \in \mathcal{T}} P(t|s^*_{so}, s^*_a, c^*_{so}, c^*_a, p) \quad (5)$$

各タスクにおける確率 P は対数線型モデルとした．

3.2 情報発信構成の同時同定法

同時同定法においては，情報発信構成の構成要素を確率変数とする条件付き確率場を考えて，確率を最大化する構成要素の組み合わせを選択することにより，情報発信構成の構成要素を同時に同定する．本稿では，図 4 に示すような依存関係をもった条件付き確率場を仮定する．

同時同定の問題は次のように定式化される．Web ページ X が与えられたとき，因子ベクトル ϕ^k を用いて，情報発信構成 $Y =< t, s_{so}, s_a, c_{so}, c_a >$ の条件付き確率が次式で与えられるものとする．

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_k \Lambda^k \phi^k\right) \tag{6}$$

ここで，$Z(X)$ は分配関数，Λ^k はパラメータである．因子ベクトルは 1 つの確率変数を含むもの 5 つと隣接する 2 つの確率変数を含むもの 4 つの合計 9 個を考える．推論は forwad-backward 法，デコードは Viterbi 法により実現できる．訓練データは $T = \{X_i, p(s^i_{so}|X_i), p(s^i_a|X_i), t^i, c^i_{so}, c^i_a\}$ という形で与えられる．c_{so}，c_a および t はラベルとして与えられるが，s_{so} および s_a については確率分布として与えられる．s_{so} と s_a を確率分布として与えるのは，ページ毎に情報発信者の候補が異なることから事前にラベルが列挙できないことと，表記の揺れにより複数の候補が正解となりうる（「株式会社 ABC 商事」と「ABC 商事」など）ためである．このような形で訓練データが与えられたとき，モデルの学習は次の損失関数を最小化するパラメータを求める問題となる．

$$\begin{aligned}\mathcal{L} \quad = \quad & \sum_i^{|T|} \{\mathrm{KL}(\tilde{p}(s^i_{so}|X_i))||p(s^i_{so}|X_i) + \\ & \mathrm{KL}(\tilde{p}(s^i_a|X_i))||p(s^i_a|X_i) + \\ & \log p(c^i_{so}|X_i) + \log p(c^i_a|X_i) + \log p(t^i|X_i)\} \\ & + \Lambda\Lambda^t \end{aligned} \tag{7}$$

訓練データが分布として与えられる変数については Kullback-Leibler 情報量を，ラベルとして与えられる変数については対数尤度を計算している．また正則項としてパラメータの L2 ノルムを用いた．

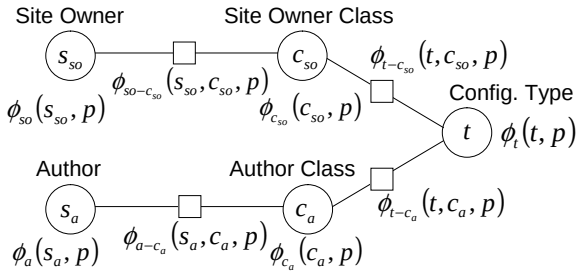

図 4: 情報発信構成の同時同定モデル．

4 実験

4.1 方法

評価用のデータとして情報信頼性分析評価用データ [2] の 20 トピック，各トピック約 100 ページの合計約 2000 ページのうち，含まれる情報発信者がたかだか 2 であるようなページ 1740 ページを対象とした．このデータには人手により情報発信構成が与えられているが，情報発信者候補 (s_{so} および s_a) の分布については，自動抽出された候補このデータに基づき，訓練および評価事例を次のように自動的に作成した．ルールに基づいて抽出*された情報発信者候補について，情報発信構成から取り出されたサイト運営者あるいは著者の名前との一致度に基づきスコアを与え，スコアの合計が 1 となるように正規化してして正解分布とした．こうして得られたデータを使いトピックをまたがない形で 5 分割の交差検定で各種法の同定精度の評価をおこなった．

4.2 結果

表 1 に評価結果を示す．表では，逐次法，同時法のそれぞれについて，情報発信構成の各要素個別の同定精度と，情報発信構成全体の精度を示している．ここで，オラクルとは逐次法で用いた個別変数のモデルについて，前段階の同定結果が正解であるような入力を与えたときのモデルの同定精度である．実際に逐次法を適用する際には，前段階で誤るケースもあるので，一般にオラクルの場合よりも精度は下がる．

評価の結果，要素毎では情報発信タイプを除いて逐次法の同定精度が勝り，情報発信構成全体についても逐次法が優位であった．

*情報発信者候補の抽出法の詳細については [6] を参照されたい．

表 1: 逐次同定法および同時同定法による情報発信構成の同定精度

要素	オラクル	逐次法	同時法
サイト運営者	-	**0.716**	0.692
著者	-	**0.596**	0.576
クラス（サイト運営者）	0.702	**0.652**	0.576
クラス（著者）	0.612	**0.509**	0.485
情報発信タイプ	0.926	0.696	**0.708**
全体	-	**0.309**	0.264

4.3 考察

著者らは実験前には，次の点で同時法が逐次法に比べて高い同定精度を示すと予想していた．すなわち，逐次法では途中の段階でモデルが一旦誤ってしまうと後の段階で修正が効かないのに対して，同時法では変数間の依存関係を利用することにより，逐次法では誤ってしまうような事例でも正しい同定結果を得られるのではないかという予想である．しかしながら，実験は予想とは逆の結果を示した．

同時法の同定精度が低かった原因としては，同時法で用いた依存構造が適切でない，当初想定したような間違いを正すことのできるような依存関係がそもそも存在しない，といったことが考えられる．この点について，他の依存構造に基づくモデルについても検討の余地がある．

その他の原因として，サイト運営者や著者といった情報発信者の候補が，現在の候補抽出手法では平均的に数 10，多いときには 100 を超える候補を抽出してしまい，モデルは非常に多くの候補の中から正解を見つけなければならないことが考えられる．このような状況で，逐次法の情報発信者同定モデルは各変数を個別に同定できるのに対して，同時法では他の変数も同時に同定する必要があり，非常に困難な問題を解いていることになっている．そのために逐次法が同時法に比べてより安定的に高い精度で同定できている可能性がある．今後，情報発信候補の抽出について [7] で述べられているような手法を利用することにより，精度良く少数の候補に絞り込まれている状態にしてから情報発信構成を同定した場合に，逐次法と同時法がそれぞれどのような精度を示すか評価することも検討したい．

5 おわりに

本研究では，Web 上の情報の信頼性を分析するという文脈において，Web ページの情報発信者に関する情報を情報発信構成として記述することを提案している．本稿では，情報発信者がたかだか 2 である場合に限られるものの，初めて情報発信構成全体を同定する手法を提案し，同定精度に関する評価結果について報告した．

同時法は変数間の依存関係を利用できることから，逐次法に比べて同定精度が高くなると実験前には予想したが，実験は予想とは逆の結果を示した．同時法で仮定した依存構造の問題と，情報発信者候補が多すぎる問題が原因として考えられ，これらの点について検討の余地が残る．

情報発信者候補の絞込による同定精度の向上，および情報発信者が 3 つ以上の場合への対応は今後の課題である．

参考文献

[1] D. Kawahara, S. Kurohashi, and K. Inui. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, pp. 393–397. IEEE, 2008.

[2] 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. *Internet Research*, 18(2):155–164, 2008.

[3] T. Nakagawa, T. Kawada, K. Inui, and S. Kurohashi. Extracting subjective and objective evaluative expressions from the web. In *Proceedings of the Second International Symposium on Universal Communication*, pp. 251–258. IEEE, 2008.

[4] 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.

[5] 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.

[6] 加藤, 河原, 乾, 黒橋, 柴田. Web ページの情報発信者の同定. 人工知能学会誌論文誌, 25(1):90–103, 2010.

[7] 船山, 渋田, 柴田, 黒橋. Web ページの構造解析とメタデータ候補の抽出. 言語処理学会第 16 回年次大会論文集, 2010.

[8] 黒橋. 情報の信頼性評価に関する基盤技術の研究開発. 人工知能学会誌, 23(6):783–790, 2008.