Ｗｅｂページからの情報発信者の抽出における レイアウト情報の利用

百瀬　亮　　宮崎　林太郎　　渋木　英潔　　森　辰則

横浜国立大学工学部電子情報工学科 〒240-8501 横浜市保土ヶ谷区常盤台 79-7

Email:{ryo_m,rintaro,shib,mori}@forest.eis.ynu.ac.jp

１．初めに

Web 上に存在する情報は爆発的に増加しており、これらの情報の中には、誤ったもの、出所が不確かなものが存在する。このため、利用者が信頼できる情報を取得するための技術に対する要望が高まっている。しかし、情報の信憑性を自動的に判断することは困難であると言える。そのため、我々は利用者が情報の信憑性を判断する為の支援を行う技術の実現が課題であると考える。

信憑性の判断を支援する技術の一つとして、情報発信者の抽出が挙げられる。情報発信者には二つの種類があると考えられる。一つは、サイトの製作者や管理者等の Web ページ単位の発信者である。もう一つは、掲示板における投稿者やコメント毎の記事単位の発信者である。

我々の研究における最終目標は後者の情報発信者抽出であるが、本論文では前者である、Web ページ単位の発信者の抽出を目指すものとする。これは、利用者はページ単位の発信者の信憑性が判断できなければ、記事単位の発信者の信憑性を判断できないからである。

２．関連研究

固有表現抽出で使われていた手法による情報抽出では、テキスト全文から、求める記述を抽出していた。しかし、我々は次の仮説を立てた。

１．テキスト全文からの情報の抽出よりも、ある程度の数の文列毎に、当該情報が記述されているかどうかの２値分類を行う方が簡単である。
２．テキスト全文から発信者情報を抽出するよりも、あらかじめ絞り込んだ文列から情報を抽出する方が簡単である。

この仮説に基づき、塚原らはオークションサイトの出品情報文書からの商品の属性-属性値抽出を行っている[3]。この研究では、１段階目において属性-属性値が記述されている文の絞り込みを行い、２段階目において文中の属性-属性値の抽出を行う事により、精度の向上を図っている。

多段階の学習を行う手法としては、Kaynak ら[1]が、複数の分類器を多段接続する事により、分類精度を上昇させる手法として Cascading を提案している。本論文の手法も Cascading の考え方に類似している。1 段階目では大域的な情報を用いて大まかな分類を行い、2 段階目ではより局所的な情報を用いて詳細な抽出を行うものである。しかし、Cascading が同質の分類タスクを各段で行っているのに対し、本論文での処理は 1 段目のタスクが 2 段目のタスクの処理対象を限定する為に用いられている点が異なる。

また、Pang らも評判情報分析において、文単位での意見性を判定した後に、肯定否定の分類を行うという段階的な処理を行っている[2]。

日本語の情報発信者抽出の研究としては、加藤らの研究がある[4]。この研究では情報発信者が記述されやすい場所が Web ページのメインコンテンツの前後にあると仮定して抽出を行っている。一方、本研究の手法では、Web ページ中の情報発信者記述部分の場所をあらかじめ決めるのでなく、1 段階目の抽出において動的に絞り込む点が異なる。

また、Web ページからの情報抽出としては鶴田らが、Web ページをグリッドに切り分け、主要な情報が記述されている DOM ノードを特定する手法を提案している[5]。しかし、この研究においては、一般的な Web ページにおいてどの部分に主要な情報が記述されているかを示すウェイトマップの作成のためのグリッドを使用しており、各セル毎に、当該情報が含まれているかどうかを分類しているわけではない。

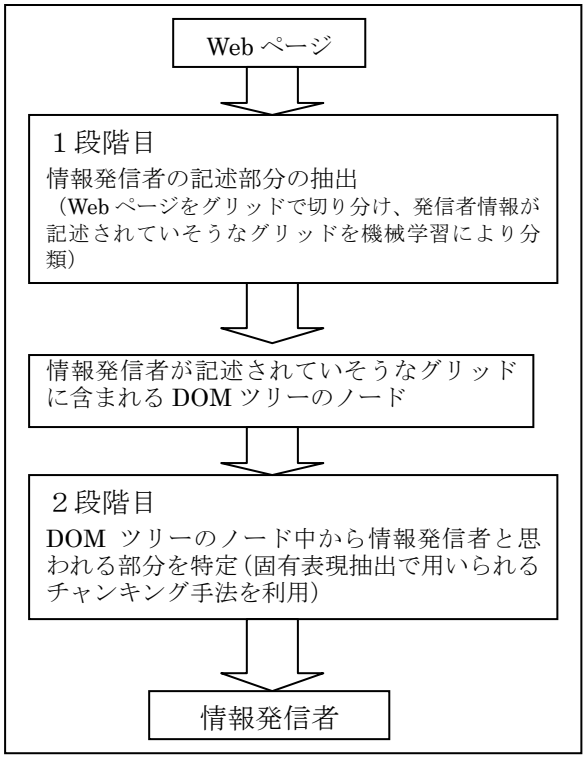

図１ システム概要

３．提案手法

本システムの概要を図 1 に示す。提案手法では 1 段階目における絞り込みを DOM ツリーのノード単位で行う。また、この際に Web ページのスタイル情報を用いる。

３．１　第１段階（DOM ノードの抽出）

１段階目の抽出では、発信者が記述されていると思われる DOM ツリーノードを、ページ中の位置情報、そのノードに含まれるテキスト等から特定する。

Web ページはその種類ごとに、情報発信者が記述されている場所が異なる。我々は、Web ページを縦横のグリッドに切り分け、このグリッドのセル毎に発信者情報が記述されているか否かを判定し、そこに含まれる DOM ノードを抽出する。各サイトにおける発信者情報の記述位置とグリッドの例を図２に示す。

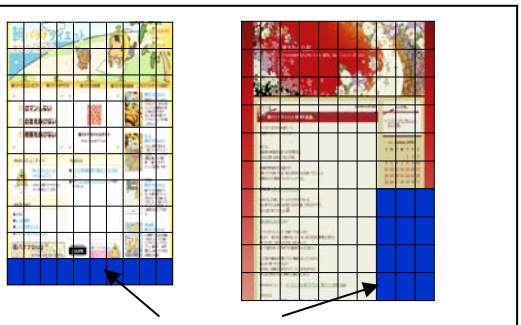

図２　情報発信者記述部分の例

３．２　第２段階（情報発信者の同定）

２段階目の抽出では、１段階目で抽出された情報発信者が記述されていると思われるＤＯＭノード中から発信者情報と思われる部分を特定する。２段階目の抽出においては、固有表現抽出で用いられるチャンク同定手法を用いる。

４．実験

本章では評価実験について述べる。評価実験は抽出の各段階を単独で行った場合の結果、連結して行った場合の結果について示す。

４．１　評価用文書

評価実験に用いたWebページは検索エンジン基盤TSUBAKI[1]を用いて収集した「無洗米」、「レーシック」について記述してあるWebページ 100 ページである。この 100 ページは全て異なるWebサイトから収集した。これは、同一発信者が記述したページを除去することにより、未知のサイトからの情報発信者の抽出を考慮したためである。

この Web ページ集合について情報発信者の記述箇所に人手で注釈付けを行ったところ、214 箇所あった。その中で、<img>タグの alt 属性に情報発信者が記述されていると注釈付けされたのは 46 個所である。また、100 ページの中で<title>部分にのみ情報発信者が記述されていると注釈付けされたものが 6 ページあった。

４．２　第１段階の結果

分類にはSVMを使用した。SVMの実装系にはTinySVM[2]を使用した。判定の単位はグリッドのセル毎とし、素性としては次のものを用いた：「グリッドセルのＸ座標」「グリッドセルのＹ座標」「グリッドセルに含まれるDOMノードに現れるHTMLタグの出現頻度」「グリッドセルに含まれるDOMノードに現れる形態素の品詞の出現頻度」「グリッドセルに含まれるDOMノードに現れる形態素の表層表現の出現頻度」「ページタイトル中の形態素の表層表現の出現頻度」「ページタイトル中の形態素の品詞の出現頻度」。

グリッドセルの座標情報の計算にはウェブブラウザの JavaScript を用いており、ウインドウサイズを 1280×1024 のディスプレイで最大化した状況で固定して、グリッドによる切り分けを行った。

実験ではグリッドを 10×10 に分割した場合と、5×5 に分割した場合を比較し、5 分割交差検定を行った。結果を表１に示す。

表１　１段階目の抽出精度

10x10		5x5	
Precision	Recall	Precision	Recall
0.21	0.52	0.48	0.68

表 1 によれば、グリッドを細かく分割しても精度の向上が見られないことが分かる。これは細分化することによりタスクが難しくなっていること、DOM ノードが描画される範囲がそれほど狭い範囲でないことが原因として考えられる。

４．３　第１段階と第２段階の結合

第 2 段階目は、系列ラベリングに基づくチャンク同定手法により行った。分類にはCRF++[3]を使用した。ラベル付与の単位は文字とし、素性には次のものを使用した：「表層文字」「品詞」「形態素原形」「文節内素性」「主辞素性」「角川類語辞典の分類番号」「活用形の原形」「Jumanにより付与された単語の代表表記」。

2 段階目への入力としては 1 段階目の分類において正解が含まれると判断されたセルに含まれるＤＯＭノードと、各ページのタイトルを与えた。1 段階目の分類ではレイアウト情報を元に分類しているため、タイトル部分については分類ができない。そのため、この実験ではタイトル部分は全て情報発信者が記述されている可能性があるものとして取り扱った。

我々の予測 2 が正しいことを確認する為に、事前に予備実験を行った。予備実験にはあらかじめ収集済みだった Web ページ 71 ページを用いた。これは、一部 4.1 節で説明した文章と同じだが、71 ページ中に同一サイトの Web ページが含まれているものである。このデータを用いて 2 段階目の手法のみによる抽出と、1 段階目で 10×10 分割を行ったグリッドにおいて情報発信者が実際に含まれるグリッドセルのみを用いて 2 段階目の抽出を行った。前者は、Web 文書の全体を抽出対象とする従来手法に相当し、後者は、1 段階目の抽出が完全に成功したことを仮定した一種の上限に相当する。表 2 に結果を示す。

この結果より、1 段階目の抽出が成功すれば、2 段階目を組み合わせた場合に情報発信者の抽出精度が向上することが確認できた。このことから、予測 2 について、その妥当性がある程度示された。

[1] http://tsubaki.ixnlp.ac.jp/index.cgi

[2] http://chasen.org/~taku/software/TinySVM

[3] http://crfpp.sourceforge.net/

表２．２段階目の抽出の精度

	Precdision	Recall
1 段階目の絞り込みを行わない従来手法	0.53	0.47
1 段階目が成功したと仮定した場合	0.84	0.47

そこで、4.1 節で説明したデータを用いて、本実験を行った。なお、本実験では２段階目の抽出における学習データについても２段階抽出に適合した学習データの作成を行った。２段階目の抽出における学習データとしてはページ全部を学習データとするのが普通だが、１段階目で絞り込みを行う点を考慮すると、文書をより狭い範囲に絞り込んだ状況での学習データの作成が考えられる。今回の実験では１段階目の抽出で使用したグリッドの切り分けを用いて、正解情報が含まれているグリッドセル中の DOM ノードのみを２段階目の抽出器に対する学習データとして用いる場合も検討した。本実験で比較検討した条件をまとめると以下のようになる。

1) ２段階目の抽出手法のみで抽出を行った場合。(baseline。従来の、全文書を対象とした情報抽出手法に対応。)
2) １段階目の抽出が完全に上手くいったと仮定した場合。２段階目の抽出に対する学習データは 1)と同じで、全文書。
3) １段階目の抽出が完全に上手くいったと仮定した場合。２段階目の抽出に対する学習データを、各ページをグリッド分割し、正解が含まれているセルのみとした場合。
4) １段階目の抽出をシステムの結果とした場合。２段階目の抽出に対する学習データは 1）と同じ場合。
5) １段階目の抽出をシステムの結果とした場合。２段階目の抽出に対する学習データは 3)と同じ場合。

実験結果を表 3 に示す。

表３ ２段階目の抽出の精度

	Precision	Recall
1)	0.41	0.15
2)	0.51	0.16
3)	0.55	0.17
4)	0.42	0.11
5)	0.45	0.12

表 3 の 1)と 2）を比較する、1 段目の絞り込みが成功した場合には、再現率を落とすことなく、適合率が 0.1 上昇している。このことより、1 段階目の絞り込みにより、2 段階目の抽出精度が向上するであろうとした我々の予測が妥当であったと考えられる。

また、3）を見ると 2 段階抽出に則した学習データの構成にした場合には baseline に比べて適合率が 0.14 上昇しており、適合率、再現率共に 2）よりも上昇していることが分かる。

次に、実際のシステムの動作に則して、1 段階目のシステムの出力を 2 段階目の入力とした場合の結果について述べる。ここで、4）の精度の上限値は 2)の結果、5)の精度の上限値は 3)の結果であることに注意されたい。4)、5)ともにわずかであるが、1)の場合に比べて適合率が上昇している。これは 1 段階目の抽出による絞り込みに効果があったことを示している。しかし、再現率については 4)、5)ともに低下している。この原因はいずれの場合も、1 段階目の抽出における取りこぼしであり、1 段階目抽出の再現率が低いことが原因として挙げられる。

また、2)と 3)を比較した場合と、4)と 5)を比較した場合の両者をみると、学習データを絞り込んだ方が適合率、再現率ともに上昇している。このことから、学習データの絞り込みは効果があると考えられる。

次に、ページ毎に情報発信者が最低一つは抽出できていたかどうかを調べた。本研究の目的は Web ページの情報発信者の提示であるから、全てを正確に抽出せずとも、一つだけ正しく抽出されるのでもよい。一方で、1 ページ内に複数個所情報発信者が記述されている場合があり、特定のページ中の情報発信者のみが抽出されやすくなっている場合を考えたからである。表 4 に結果を示す。

表４．２段階目の抽出の精度

	取得できたページ数
1)	25
2)	23
3)	26
4)	16
5)	19

ページ単位で情報発信者が抽出できたかどうかには再現率が重要となる。この点では 2 段階目の抽出における再現率は不十分である。また、1 段階目の絞り込みにおいて、誤って除去されてしまったページの影響も大きくなっている。

４．４ 失敗分析

まず、4.3 節の実験において、条件を変化させた時の抽出誤りの変化について更に細かく調べてみる。条件の変化により抽出に成功するようになった部分と失敗するようになった部分の数を求めた結果を表 5 に示す。

表５．２段階目の抽出の精度の変化

	抽出しなくなった部分		抽出するようになった部分	
	正解を抽出できなくなった	不正解を抽出しないようになった	不正解部分を抽出するようになった	正解を抽出するようになった
1)⇒2)	8	20	0	0
1)⇒4)	18	19	5	3
2)⇒3)	2	5	6	8
4)⇒5)	3	5	8	7

この結果を見ると、1 段階目の抽出による絞り込みが不要部分を抽出しなくする方向に働いることがわかる。一方、1 段階目の抽出を行ったことによ

り抽出できなくなる正解は、いずれの場合も、1 段階目の絞り込みで除去されてしまった部分である。

また、2 段階目の抽出における学習データの絞り込みはより積極的に抽出を行う方向に働いていることがわかる。これは、学習データが正解データ周辺に絞り込まれたために、負例が少なくなったためと考えられる。

さらに、各条件において、抽出すべきでない部分をシステムが情報発信者として判断している部分について調べた。いずれの場合も企業名や個人名などの情報発信者となる可能性がある部分ではあるが、そのページの情報発信者ではない部分を抽出するという誤りであった。このことから、ページ中に明に情報発信者であると記述してあるページと、人間がみて、経験から情報発信者であるということがわかるページがあるのではないかと考えた。そこで、実験に用いたデータ中で、情報発信者であることが明に書かれているページを調べた。

その結果、100 ページ中 54 ページが情報発信者であると明に記述されていないページであった。このようなページについては情報発信者の候補からどれを正解とするかといった判断が難しくなっている。

上記の調査を行った中で判明した、人間が行う判断の基準としては、以下のものがあった。

- 他に情報発信者が記述されていない
- 情報発信者がページタイトルに記述されていることから、その発信者が記述したページであると推測
- 住所などの連絡先についての記述が存在する情報発信者
- 大きく書かれている、何度も書かれている等の強調がされている

今後の課題として、情報発信者であることが明に記述されているページとそうでないページについて異なる処理を用意することが挙げられる。

また、情報発信者であることが明に書かれているページにも以下のようなものが存在した。

- 画像中に文字として情報発信者であることが記述されている
- 本文中に記述があり、細かく読まなければ分からない

特に、画像中の文字として情報発信者が記述されている場合には、必ずしも alt 属性に同一の内容が記述されているとは限らず、システムの処理としては情報発信者であることが明に記述されていない場合と同様の処理が必要になると考える。

表６ 1 段階目の結果（素性変更後）

10x10		5x5	
Precision	Recall	Precision	Recall
0.66	0.73	0.73	0.72

さらに、1 段階目の抽出による絞り込みで誤って正しいセルを除去してしまっている点について、改善できないかと考え、1 段階目の機械学習の素性のうち、出現頻度の値そのものを素性値としていたものについて、これを出現するか否かの二値の素性値に変更してみた。結果を表 6 に示す。

10×10、5×5 両者において大幅な数値の上昇が見られた。これは出現頻度を素性値にした場合、出現の有無による二値の素性値の場合に比べて、取りうる値の範囲が広がるため、学習事例数が多くない状況においては、類似事例の集約が適切に行われなかった為と考えられる。

５． まとめ

本論文では Web ページ中からの情報発信者抽出における 2 段階抽出手法を提案した。1 段階目の抽出においてはスタイル情報を利用した絞り込みを行った。2 段階目の抽出では学習データの絞り込みを行い精度の調査を行った。実験の結果によれば、データ中に同一サイトのページを含むばあい、含まない場合いずれも、1 段階目の抽出が成功した場合には適合率、再現率ともに、精度が大きく上昇する事が判明した。また、1 段階目の精度が十分でない場合でも、適合率については精度の向上が確認できた。

６．謝辞

本研究の一部は、独立行政法人情報通信機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」プロジェクトの成果である。

参考文献

[1] Cenk Kaynak and Ethem Alpaydin, “Multistage Cascading of Multiple Classifiers: One Man’s Noise is Another Man’s Data”, Proceedings of the 17th International Conference Of Machine Learning, 2000.

[2] Bo Pang and Lillian Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,” Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004.

[3] 塚原裕常, 宮崎林太郎, 西村純, 前田直人, 森辰則, 小林寛之, 石川雄介, 田中裕也, 翁松齢. ネットオークションの出品情報文書からの 2 段階属性抽出. 言語処理学会第 15 回年次大会発表論文集, pp. 400--403, (2009).

[4] Yoshikiyo Kato and Daisuke Kawahara and Kentaro Inui and Sadao Kurohasi and Tomohide Shibata, “Extracting the Author of Web pages” , Proceedings of the Second Workshop on Information Credibility on the Web, 2008.

[5] Masanobu Tsuruta and Hiroyuki Sakai and Shigeru Masuyama, “An Informative DOM Subtree Identification Method from Web Pages in Unfamiliar Web Sites”, The Institute of Electronics, Information and Communication Engineers (IEICE) Transactions Information and System, vol.E91-D, no.4, 2008