ブログページからのウェブサイト情報・作成者情報の抽出

堀 達也　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1310067,kshirai}@jaist.ac.jp

1 はじめに

我々はウェブ上でさまざまな情報を検索することができるが，ときには正しくない情報が公開されていることもある．そのため，ウェブ検索の際，ユーザは検索された情報が正しいかどうかを判断する必要がある．このとき，情報の正しさを判断する助けになるのが，ウェブサイトに関する情報や，そのウェブサイトの作成者に関する情報である．例えば，病気について調べたいときには，ウェブ検索でヒットしたウェブサイトが病院の正式なホームページであるとわかれば，そのサイトの情報は信頼性が高いと判断できる．同様に，検索によってブログ記事がヒットしたとき，そのブログの書き手が医者であることがわかれば，その内容を信用できる．

本研究では，ウェブページからウェブサイト情報や作成者情報を抽出することを目的とする．ここで,「ウェブサイト情報」(以下，単にサイト情報と呼ぶ) とはウェブサイトやブログの内容を説明したテキスト,「作成者情報」とはウェブページの作成者のプロフィールについて書かれたテキストと定義する．また，ウェブサイトや作成者の情報が記述された別ページへのリンクが存在するときは，そのリンクを抽出する．将来的には，抽出したサイト情報や作成者情報は，ユーザがウェブページの信頼性を判断する補助情報として，ウェブ検索エンジンで検索結果とともに掲示することを想定している．なお，ウェブには様々なサイトが存在するが，ブログは形式がある程度決まっているため，サイト情報や作成者情報の抽出が比較的容易であると考えられる．そのため，本論文では手始めにブログページを対象としてサイト情報・作成者情報の抽出を試みる．

2 関連研究

百瀬らは，ウェブページのレイアウト情報を利用して，二段階の手続きで情報発信者を抽出する手法を提案した [1]．第一段階では，ウェブページをグリッドに分割し，グリッドのセルごとに発信者情報が含ま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．第二段階では，抽出された DOM ノードの中から情報発信者と思われるテキストを特定する．

Kato らは，ウェブページの情報発信者情報を抽出するためのサブタスクとして，情報発信者名の抽出を試みた [2]．まず，テキストから情報発信者名を抽出する条件 (人名接尾辞が存在するかなど) や DOM(Document Object Model) の木構造におけるの深さなどを手掛かりに情報発信者候補を抽出する．それらを文書構造の特性 (HTML タグ名など) や言語特性 (人名や組織名を構成する品詞列など) という観点からランク付けし，その上位 k 番目までを情報発信者とする．

Giuffrida らは，PostScript で記述された科学論文からタイトル，著者，所属，著者と所属の関係，目次を抽出する手法を提案した [3]．テキストの位置やフォントなどの空間特性ならびに視覚特性を利用してメタデータを抽出する．実験の結果，情報抽出の正解率は，タイトルが 92%，著者が 87%，所属が 75%，著者と所属の関係が 71%，目次が 76%であった．タイトルや著者は書き方がある程度決まっているため，他のメタデータより抽出しやすいことが確認された．

これらの先行研究は情報発信者名や著者名を抽出対象としている．一方，本研究では作成者の名前だけでなく，ウェブサイトの説明文 (サイト情報) や，作成者に関する年齢，性別，職業，自己紹介文など (作成者情報) をウェブページから取得する点に特徴がある．また，HTML ファイルにおける DOM の構造やテキストを素性とした機械学習によって情報抽出を行う．

3 提案手法

3.1 概要

ブログページからのサイト情報や作成者情報の抽出は，HTML ファイルにおける DOM の個々のノードに対し，そのノードがサイト情報や作成者情報を含むか否かを判定することで実現する．サイト情報や作成者情報がタグ付けされたブログページの集合を用意し，上記の判定を行う分類器を教師あり機械学習によって獲得する．機械学習アルゴリズムは Support Vector Machine (SVM) を用いた．

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

3.2 分類クラス

DOM ノードの分類クラスを以下のように定義する．

site サイト情報を含む DOM ノード

person 作成者情報を含む DOM ノード

site-link, person-link サイト情報，作成者情報が別ページに記述されているとき，それへのリンクを含む DOM ノード

site-part, person-part テキストの一部のみがサイト情報，作成者情報に該当する DOM ノード

site-image, person-image サイト情報，作成者情報を含むが，テキストではなく画像によって表示している DOM ノード

other サイト情報や作成者情報を含まないノード

例として，site および person に分類される DOM ノードの領域を図 1 に示す．

図 1: site および person を含むブログページの例

3.3 素性

機械学習に用いる素性は node+infor という形式で表現する．node は，素性を取り出す DOM ノードを表わす．本研究では，node は判定対象の DOM ノード (N_t)，N_t の親ノード (N_p)，N_t の 1 つ前に出現する兄弟ノード (N_s)，N_t の親の 1 つ前に出現する兄弟ノード (N_{ps}) のいずれかとする．すなわち，判定対象のノードだけでなくその周辺のノードから得られる情報も素性として利用する．一方，infor はサイト情報や作成者情報の存在の有無を判定する手がかりとなる情報を表わす．SVM の学習に用いる素性ベクトルの重みは，node に infor に該当する情報が存在すれば 1，それ以外は 0 とする．

本研究における infor の一覧を以下に示す．ただし，(N_t のみ) と注記されたものは例外的に node が N_t の場合のみ素性とすることを表わす．

DOM ノードのタグ名　HTML タグは情報抽出の有力な手がかりとなる．

id, class の属性値　id=”title” や class=”profile” のように，id や class の属性値にはサイト情報や作成者情報を示唆するキーワードが含まれていることがあるため，素性とする．属性値にスペース，ハイフン，アンダーバーが含まれている場合，これらで属性値を区切る．例えば，id=”title-top” という属性値からは ‘title’, ‘top’ の 2 つの素性を得る．

テキスト長 (N_t のみ)　DOM ノードが支配するテキストの長さを l とし，l が $[1,20]$, $[11,30]$, ..., $[181,200]$ の範囲にあるとき，もしくは $l = 0$, $l > 200$ のときに重みを 1 とする素性を導入した．サイト情報や作成者情報のテキストは短いものが多く，一方ブログ本文のテキストは長いと考えられるため，テキスト長は両者を区別するために有効である．

自立語　DOM ノードが支配するテキストに含まれる自立語を素性とする．これは，サイト情報には「ブログ」，作成者情報には「年齢」「性別」などのキーワードが頻出するといった傾向を学習するためである．ただし，テキストが長い場合に素性数が多くなり，過学習を引き起こすことを避けるため，N_t のノードからは先頭から 20 番目まで，それ以外のノードからは先頭から 3 番目までに出現する語のみを素性とする．

タイトル (N_t のみ)　N_s もしくは N_{ps} が支配するテキストがそのページの ⟨title⟩ タグの内容 (ブログページのタイトル) と一致しているとき重みを 1 とする素性．この素性は，ブログタイトルと同一のテキストの近くにサイト情報が出現しやすいという観察に基づいて設計した．例えば，図 2 は図 1 に示したブログの DOM の一部であるが，⟨h2⟩ タグが判定対象の DOM ノードのとき，その 1 つ前の兄弟ノード (⟨h1⟩ タグ) が支配するテキスト「しゃかしゃか 3 人娘との毎日」はこのページの ⟨title⟩ タグと一致しており，タイトル素性の重みが 1 となる．

サイト情報を示唆するキーワード (N_t のみ)　上述のタイトル素性の重みが 1 であり，かつテキストの文末が「です」「ます」「ブログ」「日記」であるとき重みを 1 とする素性．

サイト情報へのリンクを示唆するキーワード　ノードが支配するテキストが「このブログについて」「〜

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

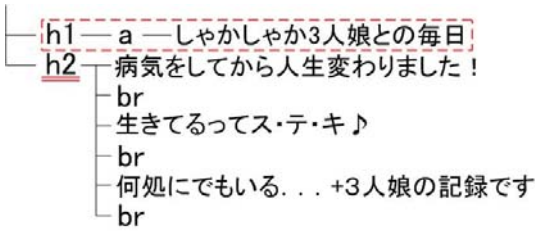

図 2: タイトル素性

ブログとは」「ABOUT」というキーワードを含むときに重みを 1 とする素性.

3.4 負例のフィルタリング

本手法の問題設定では，正例 (サイト情報や作成者情報を含む DOM ノード) よりも負例 (情報を含まない DOM ノード) の方が圧倒的に多い．実際，4 節の表 1 に示すように，実験に使用したデータにおける負例の占める割合は 99%である．訓練データにおける分類クラスの数に極端な偏りがあるのは望ましくない．

ウェブページは，そのページの主な内容を記述するコンテンツ領域と，サイト内リンク，目次，広告などを表示する非コンテンツ領域に分けることができる．サイト情報や作成者情報は非コンテンツ領域に配置されると考えられる．そこで，ウェブページのコンテンツ領域を自動的に検出し，それに含まれる DOM ノードは全て負例とみなして訓練およびテストデータから除外することにより，正例数と負例数のバランスを是正する．この処理を「負例のフィルタリング」と呼ぶ．本研究では，コンテンツ領域の検出アルゴリズムは Kato らの手法 [2] を用いた.

4 評価

4.1 実験データ

まず，実験データとするブログページを収集した．ウェブには Yahoo!，goo，FC2 などのブログサービスが存在するが，様々なブログサービスのブログを収集するために,「人気ブログランキング」というサイト [1] からランキング上位 500 件のブログのトップページを取得した．次に，これらのブログページに対し，サイト情報と作成者情報を人手でタグ付けした．表 1 に分類クラス毎にタグ付けした DOM ノード数を示す．今回の実験では，site-part と person-part は数が少なかったため，それぞれ site もしくは person と同じとみなした．また，site-image や person-image は数も少なく，また画像で表示された情報を抽出することは難しいことから，今回の実験では抽出の対象外とし，other と同じであるとみなした．タグ付けされた 500 件のブログページのうち，50 件をテストデータ，450 件を訓練データとして実験を行った.

[1] http://blog.with2.net/

表 1: 分類クラス別の DOM ノード数

site	253	person	205
site-link	14	person-link	171
site-part*	1	person-part*	5
site-image*	8	person-image*	2
		other	668370

4.2 ベースライン

提案手法との比較のために，以下のルールにしたがって DOM ノードを分類するベースラインシステムを構築した.

1. N_s または N_{ps} のテキストがブログのタイトルと一致するとき，site と判定する.
2. DOM ノードが支配するテキストが「について」「とは」という文字列を含み，かつタグが ⟨a⟩ であるとき，site-link と判定する.
3. N_s または N_{ps} のテキストが「プロフィール」「profile」であるとき，person と判定する.
4. N_t のテキストが「プロフィール」「profile」であり，かつタグが ⟨a⟩ であるとき，person-link と判定する.
5. それ以外は other と判定する.

4.3 実験結果

実験結果を表 2 に示す．M_{bl} はベースライン，M_{svm} は SVM で機械学習した分類器を用いたシステム，M_{fil} は 3.4 項で述べた負例のフィルタリングを適用したシステムを表わす．DOM ノードを表 1 の * のついていない 5 つのクラスに分類したとき，other 以外のクラスに対する精度，再現率，F 値を示した．また，参考のため，以下の 3 つの二値分類タスクの結果も掲載した.

- 簡易抽出タスク
  サイト情報，作成者情報を区別せず，いずれかを抽出するタスク．シテスムは other かそれ以外かを判定する.
- サイト情報抽出タスク
  site もしくは site-link を正例とみなし，正例か否かの二値分類を行うタスク.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 実験結果

	精度			再現率			F 値		
	M_{bl}	M_{svm}	M_{fil}	M_{bl}	M_{svm}	M_{fil}	M_{bl}	M_{svm}	M_{fil}
site	0.358	0.695	0.650	0.774	0.516	0.419	0.490	0.593	0.510
site-link	0.250	—	—	0.667	—	—	0.364	—	—
person	0.154	0.905	1.000	0.679	0.679	0.643	0.252	0.776	0.783
person-link	0.733	0.700	0.733	0.647	0.824	0.647	0.688	0.757	0.688
簡易	0.272	0.710	0.657	0.734	0.620	0.557	0.397	0.662	0.603
サイト情報	0.360	0.667	0.579	0.794	0.353	0.324	0.495	0.461	0.415
作成者情報	0.217	0.780	0.771	0.667	0.711	0.600	0.328	0.744	0.675

- 作成者情報抽出タスク
  person もしくは person-link を正例とみなし，正例か否かの二値分類を行うタスク．

4.4 考察

提案手法 (M_{svm} と M_{fil}) によって site-link と判定された DOM ノードは存在せず，抽出に完全に失敗している．これは訓練データにおいて site-link とタグ付けされた事例が少ないためと考えられる．

M_{bl} と M_{svm} を比較する．site については，M_{svm} は再現率では M_{bl} を下回るものの，精度は大きく上回り，F 値もやや高い．M_{bl} では，サイト情報はブログタイトルの周辺から取得されており，このときの再現率が比較的高いことから，多くのサイト情報が実際にブログタイトルの周辺に出現していることがわかる．一方，提案手法でも，この情報は 3.3 項で述べたタイトル素性に反映されているが，再現率が低いことから，この素性だけではサイト情報が抽出されないことが多い．したがって，サイト情報に頻出する素性を新たに導入する必要があるだろう．一方，person や person-link については，M_{svm} は M_{bl} を大きく上回ることから，作成者情報の抽出には機械学習に基づく手法が適していると言える．抽出に失敗した事例を分析したところ，ブログのプロフィール欄に家族やペットなどの紹介が書かれていることがあり，これを誤って作成者情報と判定することが多かった．

簡易抽出タスクで M_{bl} と M_{svm} を比較すると，精度では M_{svm} の方が高いが，再現率では M_{bl} の方が高い．サイト情報抽出タスクでも同様の傾向が見られるが，F 値でも M_{svm} は M_{bl} を下回った．作成者情報抽出タスクでは，全ての指標で M_{svm} は M_{bl} を大きく上回った．全般に，サイト情報よりも作成者情報の方が F 値が高いことから，後者の方が比較的抽出が容易であると言える．

M_{svm} と M_{fil} を F 値で比較すると，person については M_{fil} は M_{svm} を上回ったが，それ以外では M_{svm} の方が良い．負例のフィルタリングがあまり有効に働いていないことがわかる．今回の実験データでは，負例のフィルタリングにより，負例の数は 277014(41%) に減らされるが，誤って 76 個 (12%) の正例も除去しており，これが主に再現率の低下を招いている．

5 おわりに

本論文では，ブログページからサイト情報や作成者情報を抽出する手法について述べた．最後に今後の課題について述べる．サイト情報については，ブログタイトル近くに出現することが多いことは確認できたが，一方タイトル近くに出現する全てのテキストがサイト情報ではないため，これを識別するための素性を考案する必要がある．また，百瀬らの手法 [1] のようにレイアウト情報を素性として利用することも検討したい．負例のフィルタリングについては，サイト情報や作成者情報を含まない DOM ノードのみを削除できるように手法を洗練する必要がある．ブログ以外の一般のウェブページを対象にサイト情報や作成者情報を抽出することも試みたい．

参考文献

[1] 百瀬亮, 宮崎林太郎, 渋木英潔, 森辰則. 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, 言語処理学会第 16 回年次大会, pp.94-97, 2010.

[2] Yoshikiyo Kato, Daisuke Kawahara, Kentaro Inui, Sadao Kurohashi and Tomohide Shibata. Extracting the Author of Web Pages, Proceedings of the Second Workshop on Information Credibility on the Web, pp.35-42, 2008.

[3] Giovanni Giuffrida, Eddie C.Shek, and Jihoon Yang. Knowledge-Based Metadata Extraction from PostScript Files, Proceedings of the fifth ACM Conference on Digital Libraries (DL 2000), pp.77-84, 2000.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.