ウェブ形態情報付加によるがん情報分類精度に関する検討

木村 俊也† 中川 晋一†‡* 三角 真‡ 山岡 克式* 酒井 善則* 島津 明†

†北陸先端科学技術大学院大学情報科学研究科 〒923-1211 石川県能美市旭台 1-1

‡情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

*東京工業大学大学院理工学研究科 〒152-8500 東京都目黒区大岡山 2-12-1

E-mail: †{s-kimura, shimazu}@jaist.ac.jp, ‡{snakagaw, misumi}@nict.go.jp

*{nakagawa, yamaoka, ys}@net.ss.titech.ac.jp

1. はじめに

ウェブで提供されるがんの情報は 日々増加している. 最新のがん情報[1]を的確に得ることは延命や治療のために、手術、内服薬に匹敵する薬であるといわれている. しかし、提供されているがん情報はこの疾患の標準的な治療法が確立されていないことも原因して必ずしも良質の情報ばかりとはいえない[1].

木村・中川は胃がん、肺がん、大腸がん、子宮がん、白血病の５つのがんに関するウェブページを C-1:専門医療機関や教育機関による 研究業績などの高度な内容、C-2:個人医師や患者個人による患者指向の内容、 C-3:個人を対象としたポータルサイトや書籍の情報、C-4:個人を対象とした商品情報、C-5:検索ノイズの５種類のカテゴリに機械学習を用いて自動分類し、現在検索エンジンの提供する URL を外的基準により再評価する必要性を報告した[2][3]. この研究からがん情報はウェブページに出現する単語を素性として Naive Bayesian Classifier を用いて, 高精度（accuracy > 80%前後）で分類できることが示された. しかし、

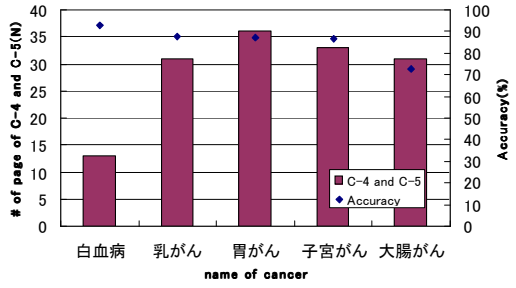

図 1： C-4 のページ数と分類精度の関係

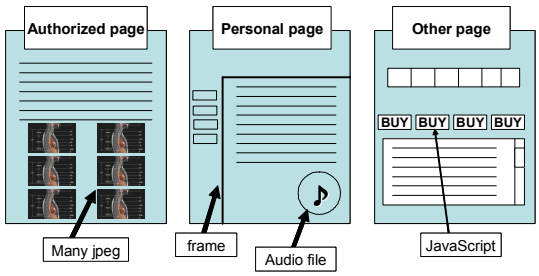

図 2： ウェブ形態の例

例えば図 1 に示すように C-4(商品情報など)のウェブページは言語モデルだけでは分類が困難であることも示唆された. 本図から C-4 のページが少ない``白血病"は分類精度が良く、C-4 のページが増えるほど、分類精度が悪くなっていることがわかる. この問題は、主に以下に示したようなウェブページが存在するために発生すると考えた.

- C-4 には商用誘導を企むページが存在し、ウェブページ上に販売を目的とした箇所と、がんの疾患を解説するためのスペースが混在しているページがあるため.
- 個人や業者ががんの疾患を説明するために公的な機関によって発信されたウェブページを参照して記述したウェブページがあるため.

この問題を解決し、がん情報をより正確に分類するため木村・中川らは、言語情報以外にイメージの総量、html ファイルの総数といったウェブページに特有に現れる分類に有効な素性(図２)を発見し、その有用性を検討した[4]. 検討の結果、今回、これらのうちウェブの形態的な素性７値等の客観的に計測可能な素性が、本稿では、この７値の素性を従来行なってきた形態素解析を用いた言語素性に加えて分類した結果、分類精度に与えた影響に関して検討する.

[1] 専門家では, ``癌”は固形癌を表す場合が多く, 白血病や肉腫などの疾患群を含めるために, 国立がんセンターではあえて``がん”とひらがなで表記する. 本研究もこれを採用する.

[1] 本研究ではがんに関するウェブ ページをがん情報と呼ぶ

表 1：実験に用いるウェブ形態素性 7 値

素性名	説明
jpeg総量	各ページ上にあるjpegイメージの総量(KB)
author文字数	authorとはmetaタグの一種でありページの作成者や所属などが表記される.
description文字数	discriptionとはmetaタグの一種でありページの要約が記述される.
JavaScriptの有無	各ページでJavaScriptが使用されているか.
ファイルの深さ	ページのドメインネームからの深さを計測したもの.
トップドメイン情報	ページのトップドメイン. 具体的にはco.jpやac.jpがある.

表 2：実験に用いる言語形態素性 6 値

素性名	説明
句点の数	各ページの半角・全角の句点をカウントしたもの.
読点の数	各ページの半角・全角の読点をカウントしたもの.
文字列総量	各ページの全ての文字列のバイト数.
行数	各ページの文書中の行数をカウントしたもの.
平均形態素数	各ページの文書中の総形態素数を行数で除算したもの.
平均文字総量	各ページの全ての文字列のバイト数を行数で除算したもの.

2. 提案手法

本研究では[4]で示された分類に有効なウェブの形態的な素性 7 値(本研究ではこの素性をウェブ形態素性と呼ぶ)に言語に関する素性を加えた結果分類精度に与える影響を考察する．以降ウェブ形態素性と言語に関する素性について説明する.

2.1. ウェブ形態に関する素性

この 7 値の素性はウェブ構造の素性２０値に対し統計的検討を行った結果選択された．詳しくは[4]を参照されたい．ここでは実験に用いる 7 値のウェブ形態を説明する．7 値の素性を表 1 に示す．専門用語数比とは、文書中に生起するすべての名詞の総頻度中の専門用語の総頻度の割合をとったものである．文書の形態素解析には松本ら[5]による Chasen + ipadic を使用した．なお、専門用語が識別できるように ipadic には木村・中川ら[6]が作成したがん専門用語集 3315 語とウェブ上から収集して作成した医学専門用語約 59533 万語[7]を追加した．専門用語比の式を(1)に示す.

$$techniq\_rate_i = \frac{\sum_{j=1} f(T_j)}{\sum_{k=1} f(W_k)} \qquad (1)$$

f(Tj)はウェブページ i において出現するすべての専門用語の頻度である．f(Wk)はウェブページ i において出現するすべての名詞と専門用語の頻度である．author と description は html タグの head 要素の子要素である meta 要素の一要素である[9]. author 要素はウェブページの作成者や所属などを記述するものであり，この文字数を素性とする．また description 要素はウェブページの内容の要約を記述するものであり,この文字数を素性とする.

2.2. 言語に関する素性

本節では、がん情報の分類に用いる言語的な素性に関して説明する．言語に関する素性は文書中に出現する名詞の頻度と，言語の計量的特徴を素性として用いる.

2.2.1. 文書中の単語に関する素性

文書中に出現する一般名詞の頻度を素性として用いる．一般名詞の頻度はウェブページ d 中に出現する名詞 t の頻度を tf(t,d)として表し､ウェブページ d における名詞 t の重みを w_t^d と考える．つまり、個々の名詞の重みを $w_t^d = tf(t,d)$ として表す．このように名詞--文書行列を作成したものを素性として用いることとした.

2.2.2. 言語形態に関する素性

本稿では，ウェブ形態素性に加え，言語の計量的特徴を用いる．これは，がん情報の書き手によってがんの疾患を解説するのに使用される，文書量や一行における文書量が違ってくるのではないかという仮説から用いることにした．本稿ではこの言語の計量的特徴の素性を言語形態素性と呼ぶことにする．過去の研究では文体の計量解析に関する研究で，文書に使われた単語の長さの分布を調べ，それが作家によって異なり，作家の特徴になることなどが示された[8]. 文体の計量解析の中でも，より視覚的かつ客観的に捉えることができる 6 値を言語形態素性として用いることにした．言語形態素性 6 値を表 2 に示す．以下，それぞれ

表３：言語形態素性の平均値と標準偏差

# of Cases	675	
	Mean	Std.Deviation
読点の数	37.27	50.38
句点の数	39.52	60.33
文字列総量	3163.69	5078.35
行数	144.25	142.20
平均形態素数	8.60	14.76
平均文字総量	28.40	48.52

の素性を簡単に説明する．句点の数とは文書中に出現する句点の総頻度である．同様に読点の数とは文書中に出現する読点の総頻度である．句点・読点ともに半角･全角を区別せずにカウントした．文字列総量は各ページの文字列部分を抽出し，文字列の総量をバイト数で表したものである．行数とは，各ページの文書中の行数をカウントしたものである．なお，行は以下の定義で区切ることにした．(ｉ)読点がある場合は読点までを 1 行とする.(ⅱ)読点が無い場合は改行までを1行とする．ただし，空行はカウントしないことにした．平均形態素数は，ページ上の文書を形態素解析し，各ページにおける総形態素数を計算する．そして，総形態素数を行数で除算し一行あたりの形態素数をあらわしたものである．平均文字総量は各ページにおける総文字列総量を行数で除算し，一行あたりの文字列総量を計算したものである．表３に本研究で用いるデータセット(データセットの詳細は 3.2 節で説明する．) から得られた言語形態素性 6 値の平均値と標準偏差を示す．

3. 評価実験

3.1. カテゴリの定義

がん情報の分類は中川ら[10]による CII(Cancer Information Index)の中から以下の 3 つのカテゴリを用いた．

1. Authorized
   学術研究機関や学会などが情報発信しているがん情報．この情報は信頼性が高いものとして提供する．
2. Personal
   闘病記や医師個人により情報発信されているがん情報．この情報は有用性が高いが，信頼性は保障できないものとして提供する．
3. Other
   広告や漢方販売に順ずるページ，またはがんに関する情報をまったく含まないページ．この情報は信頼性が低いものとして提供する．

表４： データセットの詳細

病名	Authorized	Personal	Other	Total
胃がん	20	38	41	99
肺がん	15	49	30	94
大腸がん	14	44	33	91
肝臓がん	19	26	51	96
白血病	25	34	39	98
乳がん	27	27	45	99
子宮がん	16	18	64	98
Total	136	236	303	675

3.2. 実験に用いたデータセット

データセットは検索エンジン Google を用いて，``胃がん", ``肺がん", ``大腸がん", ``肝臓がん", ``白血病", ``乳がん", ``子宮がん"の計 7 種類のがんの疾患名を個々に検索クエリとして与えた結果得られたウェブページを対象とした．それぞれの検索クエリの検索結果(通常 Google などの検索エンジンでは上限が 1000 ページ提供されるが,今回はその中で各検索クエリの上位 100 ページ(計 700 ページ)を対象とした．) 700URL に対し wget を用いて対象とするページを全量ダウンロードした．すでにページが削除されているものなどをデータセットから除外し，計 675 ページを実験に用いるデータセットとした．本データセットを医師の資格を持つものによって，3.1 節の定義に従って分類された．データセットの詳細を表４に示す．

3.3. 実験方法

ウェブ形態素性の有用性および，言語に関する素性にウェブ形態素性を加えてがん情報を分類すると分類精度にどのように影響を与えるのかを考察するための実験を行った．分類器には weka[11] による SVM を用い，評価は 10 交差検定法を用いた．SVM を用いた理由は SVM では多くの素性で学習しても過学習をしにくく，分類精度が高いためである．

3.4. 実験結果

表５に一般名詞だけを素性とした素性セット，一般名詞にウェブ形態を追加した素性セット，一般名詞に言語形態素性を追加した素性セット，それらを全て組み合わせた素性セット計 5 種類の素性セットを用いて分類した結果を示す．ウェブ形態 7 値を追加したものは，分類精度が各クラスにおいて全て向上した．特に，Authorized の Recall は一般名詞だけで分類したときよりも改善され，

表５： 分類実験の結果

一般名詞			
Category	Precision	Recall	F-Measure
Authorized	0.722	0.61	0.661
Personal	0.664	0.669	0.667
Other	0.752	0.799	0.774

一般名詞 + ウェブ形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.754	0.676	0.713
Personal	0.685	0.691	0.688
Other	0.771	0.802	0.786

一般名詞 + 言語形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.707	0.603	0.651
Personal	0.66	0.665	0.662
Other	0.751	0.795	0.772

一般名詞 + ウェブ形態素性 + 言語形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.744	0.662	0.7
Personal	0.681	0.686	0.684
Other	0.766	0.799	0.782

表６： 分類の結果(Mean of F-measure)

素性セット	mean of f-measure
一般名詞	0.70
一般名詞 + ウェブ形態素性	0.729
一般名詞 + 言語形態素性	0.695
一般名詞 + ウェブ形態素性 + 言語形態素性	0.722

表７：素性セットの素性数とモデルを作成するのに要した時間

素性セット	# of feature	Time(model)
一般名詞	12252	75 sec
一般名詞 + ウェブ形態素性	12259	84.11 sec
一般名詞 + 言語形態素性	12258	85 sec
一般名詞 + ウェブ形態素性 + 言語形態素性	12265	85.91 sec

本研究の目的を満たしたと考えられる．しかし，言語形態素性に関しては，いずれとも分類精度を下げる結果を得た．表６はそれぞれの素性セットの分類結果の F-measure の平均を示した．ウェブ形態素性は約 3 ポイントの向上を得ることができたが，言語形態素性は低下する結果となった．各素性セットの分類に用いられた素性の総数と，SVM で学習モデルを作成するに要した時間を示す．各素性の追加した数が少量であるため，学習モデルを作成する時間に与える影響は約 10 秒であり，コストが小さい．少量のコストの追加で分類精度の向上を得ることができた．

まとめ

ウェブで提供されているがん情報の質的評価を目的として、参照するウェブデータの客観情報としてのウェブ形態素性７値、言語形態素性６値を７つのがんについて計６７５例の URL データを元に検討した．一般名詞（約 1.2 万語），ウェブ形態素性，言語形態素性を用いて４つの素性セットを作成しそれぞれの URL データをベクトル化し、SVM を用いて分類し３つのカテゴリー（１：専門的、２：個人的情報発信、３：商用ならびにその他）への分類精度を検討した．分類実験を行ったところ，一般名詞の頻度とウェブ形態素性から作成した素性セットで最も良い分類精度（F-measure 73%)を得た．本法は一般名詞を用いた分類時間（675 例の closed test で total 約 75 秒）に比べて約 15 秒多かったがそれ程重い処理ではなく、実用的であると思われた．

謝　辞

本研究を行うにあたり御助言を頂いた国立がんセンター若尾文彦医長，石川ベンジャミン光一博士，情報通信研究機構久保田文人博士，ならびに関係各位に深謝する．また，本研究は情報通信研究機構運営費交付金（情報通信部門）,平成 18 年度厚生労働省がん研究助成金研究総合研究「がん情報ネットワークを利用した総合的がん対策支援の具体的方法に関する研究 」若尾班等の支援を得て行った．関係各位に深謝する．

文　献

[1] NHK. NHK スペシャル-シリーズ 日本のがん医療を問う II. http://www.nhk.or.jp/special/onair/060107.html 2006

[2] 木村，中川，三角，島津，山岡，酒井．がん情報 Web コミュニティ形成のためのコンテンツ空間の検討．DEWS2006，1b-i10．2006.

[3] 中川，木村，三角，島津，山岡，酒井．患者のためのがん情報 URL リスト適正化に関する検討．DBSJ-Letters V5 N1，pp21-24．2006.

[4] 木村，中川，三角，島津，山岡，酒井．ウェブの形態情報を用いたがん情報の分類．DEWS2007．Proceeding，2007.

[5] 松本，来内，平野，松田．形態素解析システム「茶筅」 v．2.3.3 ．奈良先端科学技術大学院大学．2003 年 8 月．

[6] 木村，中川，三角，山岡，酒井，島津．Web がん情報取得のためのがん用語辞書の作成．言語処理学会第 12 回年次大会．2006.

[7] 中川，木村，三角，島津，山岡，酒井．Web がん情報評価のための単語集合の作成と検討．DEWS2007，Proceeding，2007.

[8] 金，村上，永田，大津，山西．言語と心理の統計．pp3 － pp57．岩波書店．

[9] W3C. Technical Reports and Publications. http://www.w3c.org/TR/

[10]中川，木村，三角，島津，山岡，酒井．介入的手法によるがん情報取得適正化に関する検討．DEWS2006．1b-i9，2006.

[11]Waikato University. Weka Machin Learning Project. http://www.cs.waikato.ac.nz/ml/weak/