修 士 論 文

WWW上のがん情報の分類に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

木村 俊也

2007年3月

修 士 論 文

WWW上のがん情報の分類に関する研究

指導教官　島津明 教授

審査委員主査　島津明 教授
審査委員　白井清昭 助教授
審査委員　鳥澤健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

510030 木村 俊也

提出年月: 2007年2月

概 要

昨今インターネット技術が発達し，ウェブを介してさまざまな情報提供が行われるようになってきており，ウェブ上の医療に関する情報が日々増加している．医療患者やその家族にとってウェブは重要な情報基盤のひとつになりつつある．本研究では医療情報の中でも需要の高いがん (癌) 情報に注目して研究する．がん情報が他の医療情報に比べて盛んに流通するのは，治療法が確立されつつある糖尿病や循環器疾患に比べ，施設間での診断・治療に関する見解が標準化されておらず，診断治療にあたる医師や医療機関によって生存率が異なることが問題となっているなどの背景がある．最新のがん情報を的確に得ることは延命や治療のために，手術，内服薬に匹敵する第 3 の薬であるともいわれている．ウェブ上のがん情報に関する調査を医師とともに行った結果，検索エンジンを用いてがん情報を検索すると，医師が記述したものや個人が記述したもの（闘病記など），商用の情報などが無秩序に出力され，医学に関する専門的な知識を持たない一般人にとってはどの情報が正しいのかの判断が困難である可能性が高いことを指摘した．以上の問題を解決し，がんに関する専門知識がない一般人にも，がんの情報を正しく選択できるように支援をすることが本研究の目的である．これを実現するために，がんに関するウェブページを機械学習の手法を用いて自動的に分類する分類器を作成した．この分類器は言語情報を素性として分類精度約 80%と十分な成果を得られたが，商用のがん情報は商用誘導を企むものなどが存在し，言語の素性だけでは分類が困難である問題を示した．これを解決するために言語に関する素性に加えて，ウェブページのイメージの数や，ファイルの総量といったウェブの形態的な情報を用いて分類する手法を提案した．この手法により，言語情報だけで分類するよりも分類精度が向上することを示した．

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

昨今インターネット技術が発達し，ウェブを介してさまざまな情報提供が行われるようになってきており，ウェブ上の医療に関する情報が日々増加している．そこで本研究では医療情報の中でも特に需要が高いとされているがん [5] を一つのモデルとしてとらえ質的評価を与えることを目標とすることとした．特にできるだけ広く情報を得ることを目的とすることの多い患者やその家族にとってウェブは重要な情報基盤のひとつになりつつある．

がん情報[1]が他の医療情報に比べて盛んに流通するのは，治療法が確立されつつある糖尿病や循環器疾患に比べ，施設間での診断・治療に関する見解が標準化されておらず，診断治療にあたる医師や医療機関によって生存率が異なることが原因といわれている．がん[2]を宣告された患者や家族は新しく可能性のある治療法を検索し治癒の可能性の高い医療機関に移りたいという要求から少しでも多くの情報を必要となる．

中川・木村ら [19] によるわが国における WWW 上のがんの情報発信に関する調査により以下のことが判明した．“胃がん”,“肺がん”,“ 大腸がん”,“子宮がん”,“白血病” の５つのがんについて，わが国で発信されているこの分野のコンテンツは 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，2:個人医師や患者個人による患者指向の内容，3:個人を対象としたポータルサイトや書籍の情報，4:個人を対象とした商用情報，5:検索ノイズ，の 5 類型に分類できることが示された．また専門性の高い研究指向の類型 1 は根拠があり有用な情報を含むが，専門用語の知識のない患者にとって理解することが困難であり，間違った解釈を生むことも示された．

しかし一般的な検索エンジンを用いてがんに関する情報を検索すると，上記の 5 類型の情報が無秩序に出力され， 医学に関する専門的な知識を持たない一般人にとってはどの情報が正しいのかの判断が困難である可能性が高いことを指摘した．また，商用のがん

[1]本研究ではウェブ上のがんに関する情報を省略して “がん情報” と呼ぶことがある．

[2]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” とひらがなで表記している．本研究でもこれを採用する．

情報ページには，有用でありうるがんに関する情報が記述されているが，商用誘導を企てているページが存在するため，がんの治療法を探しているがん患者を困惑させてしまう可能性が高い．

これらの情報は人の生命に関わる重要情報であるにもかかわらず，社会財としての客観的評価を与えることが難しく，医学的根拠のない民間商用誘導なども問題になっている [5]．

1.2 研究の目的

がんに関する情報は必ずしも専門的な情報が患者のニーズに適合するわけではない．また，がんは病期や進行によっても必要な情報が異り求める情報は多種多様に存在する．このように具体的でより患者のニーズに近い情報は類型 2 の闘病記などに存在することが予想される．これを可能にするにはウェブ上のがんに関する情報を背景で述べた 5 類型に自動的に分類して提供されなければならない．

これを実現させるためには，まずがんに関する文書を正確に解析できなければならない．しかし，がんには非常に多くの専門用語が存在し，かつ治療法なども考慮すると常に用語は増加している．これらの全ての専門用語を既知とするのは困難であるため，がんに関する情報で標準的に用いられる専門用語を検討する必要がある．

本研究での分類対象はウェブ上の文書であることを考慮しなくてはならない．ウェブページは量が多く有用であるが，ウェブ上のがん情報の場合では，商用誘導を企む文書や文書がほとんどないページが存在するために，これらの問題を考慮した分類をしなければならない．

本研究では一般的に使用される検索エンジンでは無秩序に出力されるがんに関す情報を情報の発信元を外的基準として自動分類し，がんに関する専門知識がない一般人にも，がんの情報を正しく選別できるように支援をすることが本研究の目的である．

1.3　本論文の構成

本論文の構成は以下の通りである．

2 章では，医療情報のマイニングや文書分類や本研究の特色について述べる．

3 章では，わが国におけるウェブ上のがん情報流通状態に関して述べる．

4 章では，がん情報を解析するためのがん用語辞書に関して述べる．

5 章では，がん情報の分類実験を行った結果と言語空間に関する考察を述べる．

6 章では，ウェブの形態に特有に現れる素性の検索とその素性の有用性について述べる．

7 章では，ウェブの形態的な素性と言語素性を組み合わせた分類実験とその考察を述べる．

8 章では，本研究のまとめ，及び今後の展望について述べる．

第2章　関連研究

本研究はウェブ上のがんに関する情報の自動分類やマイニングを行い，がん情報の検索者にスムーズに情報を提供するシステムの開発を検討する．これまでの研究で医療情報から治療法や疾患名の抽出を試みた研究報告がされてきた．WWW の急速な発達により，医療情報の増加などの背景からウェブ上の医療情報の分類及びマイニング研究が活発に行われ始めてきた．しかし，我々の研究のように特定のドメインでのウェブ上の文書分類に関する研究はあまり報告が見られないため，オープンドメインでのウェブ上の文書分類に関する研究と，ウェブ上の医療情報のマイニングに関する研究をいくつか示し，本研究の特色を示すこととした．

2.1　ウェブ上の文書分類

落谷ら [12] による研究では，分類対象のデータセットに YAHOO!JAPAN などのインデックスサービスを用いているため一般ドメインでの分類問題となる．我々の研究ではがんに関する情報に絞っているため，特定ドメインテストであり，ドメインの違いはあるが，ウェブページを分類するという点は同様であると考えられる．

落谷らの研究では，ウェブページ中の文書を形態素解析にかけ，形態素，形態素の bigram，連語を素性としてウェブページを分類している．我々の研究でも形態素 (名詞) を素性として分類する予備実験を行ったが，ウェブ上の文書には商用誘導や他ウェブページの文書の引用したページなどが存在するために分類を誤判別してしまうものがある．本研究では，単純なテキストデータには無いウェブページ特有に現れる素性も利用して自動分類を試みる．

2.2　ウェブ上の医療情報データマイニング

ウェブ上の情報を用いた医療情報のマイニングに関する研究では長沼ら [23] の研究があげられる．長沼らは，検索エンジンを用いてウェブ上から検索者が必要としている疾患

に関するウェブページをダウンロードし，ウェブページ上の文書の内容を解析をする．解析したウェブページから知りたい項目 (症状，原因，治療方法) の候補群を作成し，検索者に提供するシステムである．長沼らによる研究は膨大に存在する WWW 上のデータから必要な箇所だけを抽出し検索者に提供するシステムであり，大変有用であると考えられる．しかし，今後ますます増加していく WWW 上のデータからこれら諸項目を抽出すると，検索者はその中から信頼できる情報を抽出することが困難になることが予想される．本研究では，ウェブ上のがん情報を情報の発信元を推定することにより情報の信頼性を付加して提供することが可能となる．

がんの専門用語の作成に関しては中川 [21] [16] らの研究があげられる．中川は国立がんセンターが提供する 53 種類の疾患解説ページから，手作業でがんに関する専門用語 3316 語を切り出した．本研究ではこの 3316 語をがん情報の解析に用いることにした．

2.3 ウェブ上の医療情報のメタデータの仕様に関する研究

Malet ら [3] はウェブ上の医療情報に関して，医療情報専用のメタデータの仕様の作成に関する研究を行っている．

第3章　わが国におけるウェブ上のがん情報

3.1　調査方法

ウェブ上に存在するがんに関する情報を獲得し，わが国におけるがん情報の流通状態を調査した．がん情報の獲得には一般的によく用いられる検索エンジンを用いる．そして，検索エンジンから得られたURLリストを用い，HTMLファイルをダウンロードし，データとして固定する．これらに対して複数人の評価者がカテゴリ分類を行いカテゴリのタグ付けをした [19] [20].

3.1.1　URLリストの固定

Yahoo! JAPANによる検索エンジンを用い，検索クエリとして次の5種類の疾患名をそれぞれ少なくとも一つの単語を含む条件(OR)で入力し，それぞれの疾患名に対して1000個のURLリストを得た.

- 胃がん，胃ガン，胃癌
- 肺がん，肺ガン，肺癌
- 子宮がん，子宮ガン，子宮癌
- 大腸がん，大腸ガン，大腸癌
- 白血病

3.1.2　HTMLファイルの固定

得られたURLリストの中で上位100位を対象として，wgetプログラムを用いてダウンロードし，対象とするhtmlファイルを固定した.

表 3.1: CII(Cancer Information Index) の定義

C-1:	Peer Review を行っていると思われるがん専門機関によるがんに関する情報. 国立がんセンターや大学機関などの専門機関によって提供されている情報.
C-2:	個人または団体による Peer Review されていないがん情報. 医師個人による情報提供，個人による闘病記，個人病院等による情報提供など，ブログやがん情報を扱った掲示板も含める.
C-3:	メディアに対する情報提供. ポータルサイト，書籍情報.
C-4:	商用目的の情報提供. 医療情報を提供していても得られた HTML の中に商品販売や商用サイトへのリンクを含むもの.
C-5:	検索ノイズ. ウェブページの文書中にがんに関する情報を含まないもの.

3.1.3 カテゴリのタグ付け

3.1.1 節で固定されたそれぞれの html ファイルを，医師の資格を持つ者 (専門的知識を持つ)，がん患者 (専門知識を持たないがある程度の知識を持つ)，学生 (がんに関する知識を持たない) の 3 名で順不動，別々に次のカテゴリ分類を行った．カテゴリは C-1 から C-5 の 5 種類から構成され，これを CII(Cancer Information Index) と呼ぶ．CII の定義を表 3.1 に示す．このカテゴリの方式は C-1 に近づくほど，専門的であり情報の信頼性が高いと考えられ，C-5 に近づくほど，専門的ではなく信頼性が低くなると考えられる．

3.2 調査結果

表 3.2 に疾患名別のカテゴリ分類の結果を示す．合計値を見るとわかるように，カテゴリによってウェブページ数のばらつきが多く，特に C-1 が少なく，C-2が多いことが特徴的である．表 3.2 を元に作成した各カテゴリにおけるウェブページ数の割合を計算したものを表 3.3 に示す．この表からもわかるように，“医師個人や患者の闘病記が多く，専門医が記述したページが少ない”．これがわが国におけるウェブ上のがん情報流通の特徴の一つであると考えられる．

各カテゴリにおけるそれぞの疾患のウェブページが占める割合を考察するために表 3.2

表 3.2: 5 種類の疾患のがん情報を分類した結果

Category	肺がん	白血病	大腸がん	胃がん	子宮がん	Total
C-1	4	12	4	0	4	24
C-2	36	60	39	38	42	215
C-3	29	13	18	26	21	107
C-4	25	6	34	27	26	118
C-5	6	7	5	9	7	34
Total	100	98	100	100	100	498

表 3.3: 各カテゴリにおけるウェブページ数の割合

Category	rate(%)
C-1	4.81
C-2	43.17
C-3	21.49
C-4	23.69
C-5	6.83

表 3.4: 5 種類の疾患のがん情報を分類した結果 (横%表)

Category	肺がん	白血病	大腸がん	胃がん	子宮がん	Total
C-1	16.67	50.00	16.67	0.00	16.67	100
C-2	16.74	27.91	18.14	17.67	19.53	100
C-3	27.10	12.15	16.82	24.30	19.63	100
C-4	21.19	5.08	28.81	22.88	22.03	100
C-5	17.65	20.59	14.71	26.47	20.59	100

表 3.5: 各疾患におけるそれぞれのカテゴリのウェブページ数の標準偏差

	肺がん	白血病	大腸がん	胃がん	子宮がん
標準偏差	12.76	20.38	14.44	13.64	13.75

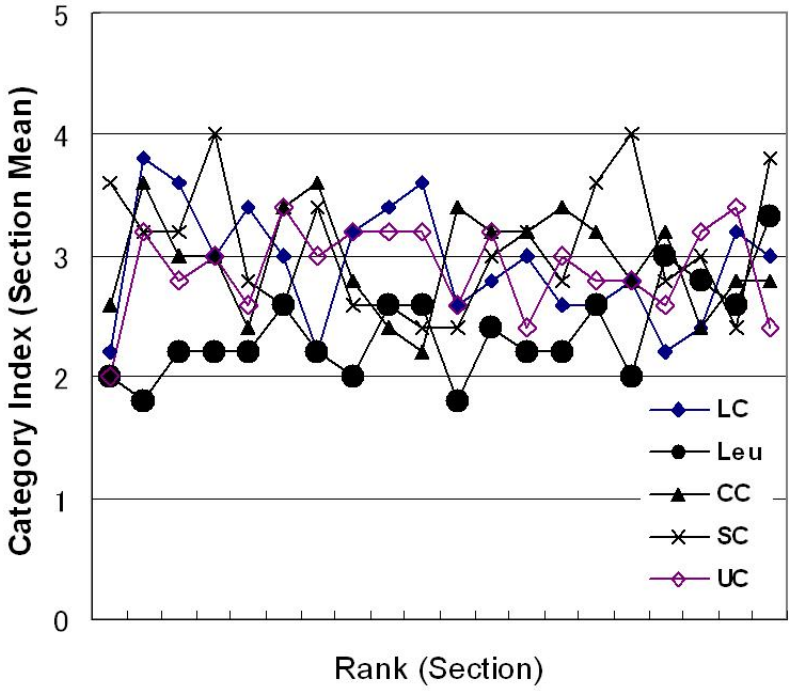

図 3.1: それぞれの疾患における検索結果のランキングの 5URL の平均値の変動.LC=肺がん，Leu=白血病, CC=大腸がん, SC=胃がん, UC=子宮がん

から表 3.4 を作成した．C-1 は白血病に多く，胃がんには無いことが特徴的であった．つまり，白血病は商用の情報が少なく，専門医が記述したものや個人が記述したものが多いことが示唆さた．このことより，疾患によって検索エンジンから提供される情報の質が違う可能性が高いことが示された．表 3.5 に，各疾患におけるそれぞれのカテゴリのウェブページ数の標準偏差を示した．この値が大きいほどカテゴリのウェブページ数のばらつきが大きく，ばらつきが小さいほどカテゴリのウェブページ数が一様であると考えられる．

図 3.1 にそれぞれの疾患での URL 検索結果の順位 (１位から１００位までにリストアップされた URL の順位ごとの 5URL ずつを区切りとした区間のカテゴリの平均値の変動) を示した．疾患別に特徴が見られ，特に大腸がん，胃がんでは上位ほどスコアが高く，白血病では順位下がっていくに従ってノイズが増加した．

以上の調査結果から次のことが明らかになった．

- わが国におけるがん情報提供状態は，専門的な情報を発するページは小数であり，専門機関よりも医師個人や患者個人によって提供される個人的情報発信が多い特徴がある．

- 胃がん，大腸がん，肺がん，子宮がん，白血病のそれぞれにおいて検索エンジンで得られた検索結果について内容を CII に従い分類した結果，それぞれの疾患により検索ランキングとノイズ比の出現率は異なっている．
- これらのことから，これらの順位付けの適正化のための中立的な機構が必要であることが示唆された．

第4章　がん用語辞書の適用

4.1　がん用語辞書の必要性

がんは，高血圧や糖尿病のように治療法の確立している疾患群とは異なり，医師にとっても特殊な用語が存在する．特に，治療方針を説明し同意を得る “インフォームドコンセント” という過程が不可欠であり，その説明のために医師も患者に対して特殊な言葉遣いをすることが多い．例えば “転移性肺がん” という用語を一般的な用語辞書で形態素解析を行うと，次のように切り出してしまう [21].

- 転移性肺がん
  - 転移
  - 性
  - 肺がん

がん情報を正しく解析する，あるいは正しく分類するためには，“転移性肺がん” は一単語として認識される必要がある．中川らによる [4] 統計的なモデルで機械的に専門用語抽出をするアルゴリズム提案されており，実装しがん用語の抽出を試み，約３万語を得たが，誤抽出が約２割ほどあり，中川により作成されたがん用語辞書 3316 語を用いることにした [16].

4.2　がん用語辞書の作成方法

がん用語辞書は，国立がんセンターのホームページにある 53 疾患のがんを解説しているページから，医師の資格者である中川によってそれぞれページにおいて手作業で専門用語を切り出された．これらを一つの語の集合とし，疾患ごとに独立して用語集合を作成する．このようにして作成された本集合の各用語の異なり語を用語辞書とした．がん用語辞書作成に用いられた 53 疾患を表 4.1 に示す [16].

表 4.1: がん用語辞書作成に用いた疾患名

悪性黒色腫	悪性リンパ腫	リンパ腫 (成人)
胃がん	陰茎がん	上咽頭がん
中咽頭がん	下咽頭がん	外陰がん
肝細胞がん	急性骨髄性白血病	急性リンパ性白血病
胸腺腫	菌状息肉症	原発不明がん
喉頭がん	骨髄異形成症候群	子宮頸部がん
子宮体部がん	子宮肉腫	絨毛性疾患
食道がん	腎盂・尿管がん	神経膠腫
腎細胞がん	膵がん	膵内分泌腫
成人 T 細胞白血病リンパ腫	精巣腫	前立腺がん
大腸がん	多発性骨髄腫	胆管がん
胆嚢がん	膣がん	中皮腫
聴神経鞘腫	軟部肉腫 (小児)	軟部肉腫 (成人)
乳がん	脳腫瘍 (小児)	脳腫瘍 (成人)
肺がん	慢性骨髄性白血病	慢性リンパ性白血病
皮膚がん	ぶどう悪性黒色腫	膀胱がん
ホジキンリンパ腫	慢性骨髄増殖性疾患	網膜芽細胞腫
卵巣がん	卵巣胚細胞腫	

表 4.2: がん用語辞書を用いた形態素解析の結果

	形態素数	未知語検出数	未知語率 (%)
がん用語辞書あり	25098	134	0.53
がん用語辞書なし	26802	265	0.99

4.3 用語辞書の妥当性の検討

辞書の妥当性を検討するため，疾患別に用語を異なり語にして加えたときの辞書内に存在する用語数について検討した結果を図 4.1 に示す．横軸はそれぞれがんの疾患であり，縦軸は得られた専門用語の総数である．疾患数が増加するにつれ，辞書の用語数も単調に増加するが，１つの疾患あたりの増分が減少する．計 53 種類の疾患の用語を全て組み合わせた結果，辞書に取り入れる用語は合計 3316 語となった．図 4.1 の疾患毎の増分の微分値をプロットしたものを図 4.2 に示す．増減があるものの単調減少であり，約 10 個の疾患で全体の単語数の約 25%を，約 20 個で約 50%を占める．次に，疾患毎に用語を加えていく過程で，疾患を 1 個加えるごとに，どれほどの用語が重複しているかを示したものを図 4.3 に示す．横軸には各疾患を，縦軸には 1 個の疾患を加えたときの重複率を示した．図 4.3 に示したように，各疾患を解説するのに用いられる専門用語は多くが重複していることがわかる．以上のことから “WWW 上でよく用いられるがん専門用語は限定されており，標準的な研究機関である国立がんセンターのウェブページで用いられている専門用語をがん専門用語辞書に収めれば，大概の専門用語はカバーできる．” という仮説を立てた．この仮説を元に本論文で作成したがん専門用語辞書を用い実際に存在するがんに関するウェブページではがん用語をどれだけカバーできるかの実験を試みた．

4.4 実験

作成した専門用語辞書を chasen（chasen-2.3.3 + ipadic-2.7.0) [24] に適用して実験した．実験方法はがん患者，完治済みのがん患者が作成した闘病記を綴ったブログページをテストデータとした．

まず，栃木がんセンターのウェブページにある，計 15 種類の臓器別診療情報の文章を形態素解析するのに本研究で作成したがん用語辞書を chasen に適用した結果得られた解析結果と適用しない場合での結果を表 4.2 に示す．

次に，それぞれのブログページに出現する専門用語を手作業で分割し，がん専門用語

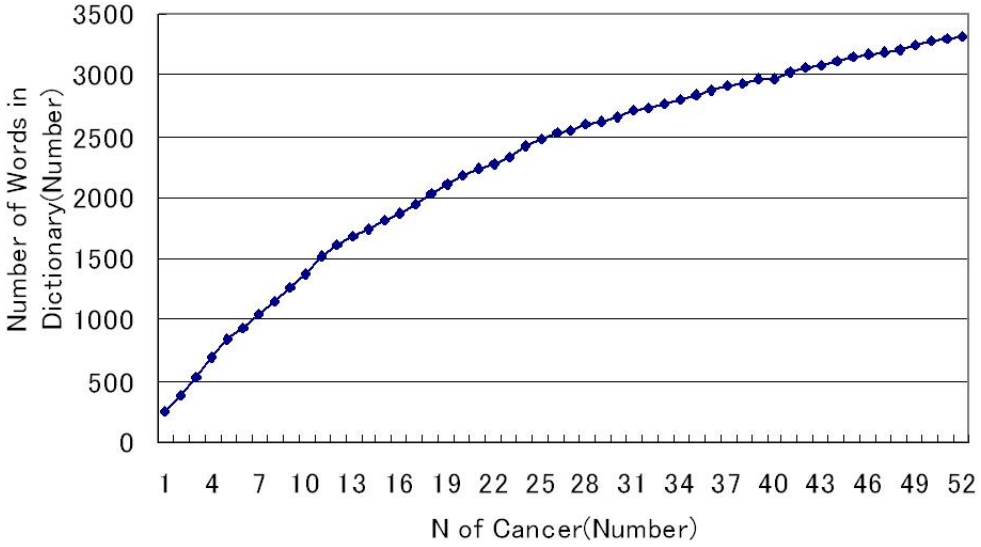

図 4.1: がん用語辞書の増加量

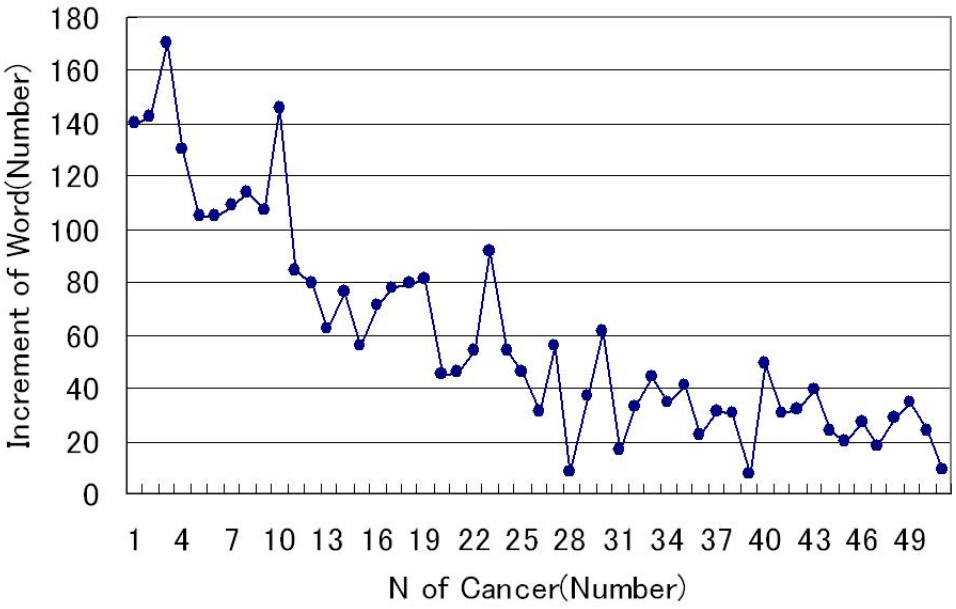

図 4.2: がん用語辞書の増加量の微分値

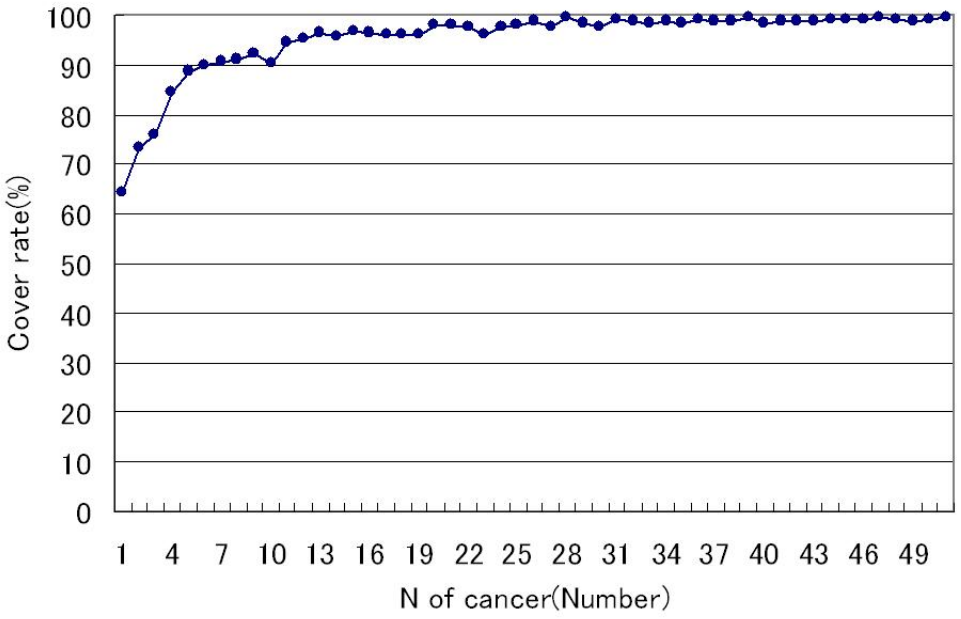

図 4.3: 各疾患での用語辞書を追加したときの用語の重複率

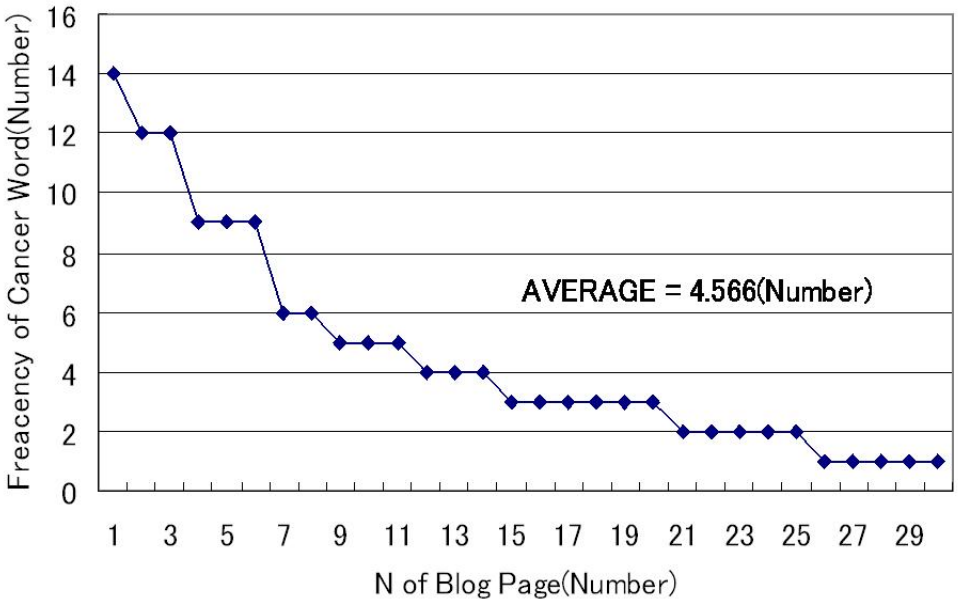

図 4.4: 闘病記に出現するがん専門用語数

辞書がどれほどカバーしているかを計測する．まず，検索エンジン goo[1] を用いて，検索クエリを “がん闘病記” として与えた結果得られたブログページをランダムに 30 ページ選出した．そしてその 30 ページに出現する専門用語を医師有資格者によって手作業で選出した．なお，得られた用語で ipadic の辞書に含まれる用語はあらかじめ削除した．その結果，各ブログページに出現した専門用語数の推移を図 4.4 に示す．なお，図 4.4 の横軸はがん専門用語の出現回数が多いブログページ順に並べた．がんに関する個人が作成したブログページに現れるがん専門用語は平均4.56 回と少ないことが示唆された．そして，個々のブログページに出現した専門用語を中川らが作成したがん専門用語集がどれほどカバーしているかを調べた結果を図 4.5 に示す．平均 65.1%の用語が辞書にある用語と重複していた．

がん専門用語辞書に含まれていなかった用語の一例を，カテゴリに分類して表 4.3 に示す．まず 1 群に現れた，“がん” の表記のずれに関して，我々はひらがなで “がん” として表記している．しかし，がんは漢字でもカタカナでも表記できる．がん専門用語辞書に，漢字で “癌”，カタカナで “ガン” を追加すると登録する用語の量が大幅に増加してしまう．

[1]検索エンジン goo, http://www.goo.ne.jp

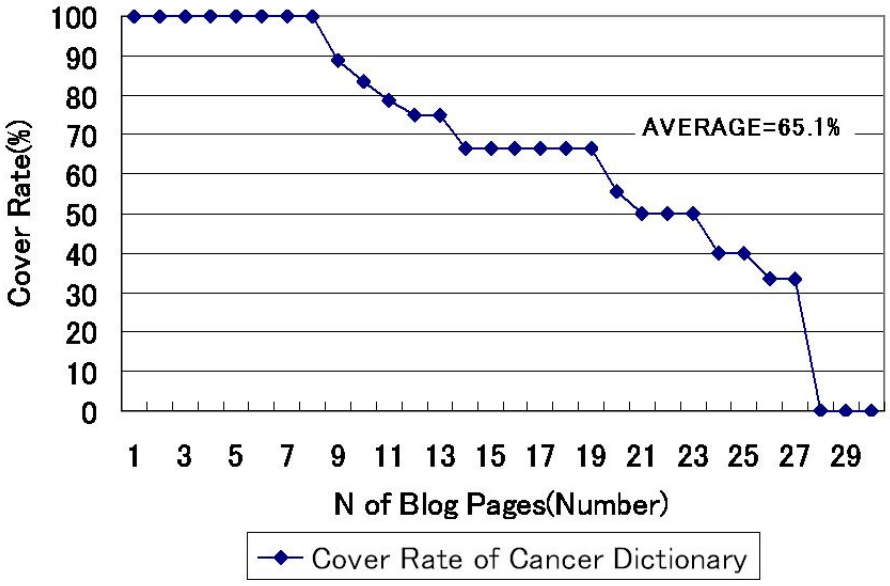

図 4.5: 作成した辞書が闘病記に出現するがん用語をカバーしている割合

表 4.3: 闘病記の文書中の用語でがん用語辞書に含まれなかった例

1 群:“がん” の表記のずれ		
抗ガン剤	子宮頸部ガン	ガン細胞

2 群:薬品名		
アレビアチン	グリオブラストーマ	ジフルカン
ハルシオン	ボルタレン	レドニン

3 群:複合語		
MRI 画像	完全麻痺	手術前投薬
麻酔前投薬		

4 群:治療法		
AdVP 療法		

これに関しては今後の検討課題にするが，がんに関する情報に対し言語的な何らかの処理をする場合は，得られた情報を一度我々が使用する言語の形式（例えば “ガン，癌” ならば “がん” にする.）に変換してから処理するといった方法を考えている．2 群の薬品に関して，薬品は種類が多く，かつ，新薬が作成される頻度も高い．よって，すべてのものを登録するわけではなく，WWW でよく使われるものの中から，危険性が低く認可されているもののみを登録する方針で考えている．これは 4 群の技術に関しても同様である．3 群に含まれる複合語に関しては依然検討中である．がん専門用語には複合語が多く存在している．表にも示したように，例えば “MRI 画像” という用語がある．我々が作成した辞書には “MRI” と “MRI 検査” が登録されているので，“MRI 画像” が未知語となることはない．しかし，複合語で成り立っている専門用語をすべて一つの形態素とするかを決定しなければならない．本研究では，一部の例外を除いて複合語を一つの形態素として登録した．例外とは，がん専門用語で特有に用いられる “原発性胃がん” や “転移性肺がん” といった “原発性” や “転移性” といった疾患の性質を意味する単語に関しては分割して形態素として適用した．

以上の検討から，中川によるがん用語辞書は本研究におけるがん情報の分類の際に文書の解析に使用するがん用語辞書として妥当であると考え，適用することにした．

第5章　言語情報を用いたがん情報の分類

ここまでの調査からウェブ上のがん情報は有用なサイトが数多く存在するが，専門医によって記述された文書は患者にとって難解であり，欲しい知識が得られない場合がある．本章では，がん情報の中でも患者に理解しやすく書かれた闘病記や患者に向けられた医師個人のページの有用性に着目した．これらのコンテンツは日記形式のものが多く断片的な記述であり，その情報を整理することによりある程度まとまった情報として提供することが可能であると考えられる．がん情報を情報の書き手によって分類する手法に関して，医師によって分類された教師データを元に各ページの文書中の言語情報を素性として学習モデルを作成し，Naive Bayesian classifier でウェブ上のがん情報を分類した．また，分類した結果，がん情報特有に表れる言語空間を調査した [15].

5.1　書き手による分類の必要性

3 章で述べたように，わが国におけるがん情報は患者による闘病記や医師個人によるがんの解説ページが多いという特徴がある．一般的に使用される検索エンジンを用いた検索結果では，医師が記述したもの，個人が記述したもの，商用のものなどが無秩序に得られるため，医学に関する専門的な知識を持たない一般人にとってはどの情報が正しいのかの判断が困難である可能性が高い．専門知識を持たないがん情報検索者の情報の選定を効率化するためには，これらの情報を整理して提供する必要がある．

5.2　カテゴリの定義

本節では，がん情報を分類するカテゴリの定義をする．カテゴリの定義は便宜のため 3 章で使用した CII の各カテゴリの呼び名をわかりやすく変更したものを用いる．本章で用いるカテゴリの定義を表 5.1 に示す.

表 5.1: カテゴリの定義

1: Authorized 学会，学術研究機関により発信された情報. Peer Review を行っていると思われる情報.
2: Personal 医師個人や患者により発信された情報. Peer Review が行われていない情報であり，闘病記，医師個人の情報を含む.
3: Media ポータル，書籍情報など.
4: Other 商用情報. 商品の宣伝など.
5: Noise 検索目的にあわないもの. ウェブページの文書中に検索クエリを含まないもの.

5.3 実装する分類器の概要

ウェブページを CII に従って自動分類するために，ベイズの定理に基づいた Naive Bayesian classifier [2] を実装した．近年，文書分類に関しては SVM などの手法のほうが多く用いられるが [8]，Naive Bayesian classifier を分類器として用いた理由は，本章ではウェブページの分類とともに，研究対象の言語空間を分析するのが目的だからである．そのため，わかりやすく実装が容易である上に分類精度も高い Naive Bayesian classifier を選択した．本章で実装する Naive Bayesian classifier の全体の処理を 2 step にわけて説明する．まず，step 1 であらかじめ正解データがついているがんに関するウェブページを教師データとして学習し，それぞれのカテゴリのトレーニングデータを作成する．そして，step 2 に処理が移り，step 1 で学習したトレーニングデータを用いて分類器の精度を測る．テストデータは検索エンジン Yahoo! JAPAN を用いてそれぞれ “胃がん”,“大腸がん”,“子宮がん”,“肺がん”,“白血病” を検索クエリとして検索した結果得られた上位 30 件を医師によって分類された結果をテストデータとした．

5.3.1 step 1

ここでは教師データを用いてトレーニングデータを作成する．つまり学習モデルを作成するプロセスである．本章で作成するトレーニングデータは，ウェブページから抽出され

た文書を教師データとし，それに対して chasen を用いて形態素解析した結果得られた名詞の頻度をカウントする．これを各カテゴリ毎に作成しトレーニングデータとする．本章では分類の素性は文書の文脈や名詞の出現箇所を考慮せずに名詞の出現頻度のみを素性とした単純なモデルで実装した．

5.3.2 step 2

step 1 でトレーニングデータを作成した後に step 2 の処理に移行する．このプロセスの処理は [7] [11] の実装を参照して作成した．step 2 では，それぞれの読み込まれたウェブページがどのカテゴリ属するかを推定する．推定するために，step 1 と同じように読み込まれたそれぞれのウェブページから文書を抽出し，その文書に対して chasen を用いて形態素解析を行い形態素に分割する．そしてそれぞれのウェブページの個々の名詞の出現頻度をカウントする．

各カテゴリを $\{c_1, c_2, \cdots, c_4\}$ とする．それぞれのウェブページを $\{d_1, d_2, \cdots, d_j\}$ とする．そして，ウェブページ d_j に出現する名詞を $\{w_1, w_2, \cdots, w_k\}$ とおき，読み込まれたウェブページ d_j に対し事後確率 $P(c_i|d_j)$ を最大化するような $\hat{c}$ を求める．$\hat{c}$ は次式で求められる．

$$\begin{aligned} \hat{c} &= argmax_{ci} P(c_i|d_j) && (5.1) \\ &= argmax_{ci} P(c_i|w_1, \cdots, w_n) && (5.2) \\ &= argmax_{ci} P(w_1, \cdots, w_n|c_i) P(c_i) && (5.3) \end{aligned}$$

そして，Naive Bayesian classifier の定義に従い，各カテゴリにおいて単語は独立に生起すると仮定し，ウェブページに割り当てられるカテゴリの推定は次の式で求める．

$$\hat{c} = argmax_{ci} P(c_i) \prod_{k=1}^{n} P(w_k|c_i) \quad (5.4)$$

(5.4) 式で，$P(c_i)$ は次式で求められる．

$$\frac{\text{トレーニングデータ中の } c_i \text{に含まれるウェブページ数}}{\text{トレーニングデータ中のすべてのウェブページ数}} \quad (5.5)$$

また，$P(w_k|c_i)$ は c_i に出現する総単語数を N_i，c_i において w_k が出現する頻度を F_{ik} とおき次のように計算する．

$$P(w_k|c_i) = \frac{F_{ik}}{N_i} \tag{5.6}$$

以上の計算がオリジナルの Naive Bayesian classifier の主な計算であるが，本論文での分類対象はがん情報であるため，新しいウェブページを読み込んだ際に教師データに現れることが無い専門用語や新語が多く出現する可能性がある．オリジナルの計算方式では確率の積をとっているため，もし一単語でも F_{ik} が 0 になると確立が 0 となってしまい，そのカテゴリには分類されなくなってしまう．そこで，この問題を解決するために [11] と同じように，予期尤度推定法で smooting を施した．これはゼロ頻度の問題を解決するために，出現する全ての単語 (名詞) の頻度に 0.5 をあらかじめ加算し，すべての単語の異なり数を V とおき，次式のように定義する．

$$P(w_k|c_i) = \frac{F_{ik} + 0.5}{N_i + 0.5V} \tag{5.7}$$

読み込まれたウェブページに出現する単語が教師データ中に存在しない，つまりゼロ頻度問題が発生したときは次式のように計算する．

$$P(w_k|c_i) = \frac{0.5}{(N_i + 0.5V)} \tag{5.8}$$

5.3.3 計算式の修正

Naive Bayesian classifier は基本的には十分なトレーニングデータが無ければ，分類精度があまり高くなく，適度な学習をすることで良い分類精度を得ることが期待できる．しかし，トレーニングデータの増加により計算なコストが高くなる．本論文が扱っているウェブ上のデータは大量に取得できることから十分なトレーニングデータを獲得することができるが，膨大なウェブページ数であるため，step 2 で説明した (5.7) 式，(5.8) 式では分母が過大化する上に，積をとっているために多くは確率が 0 になってしまう．

そこで，膨大な量のウェブページを処理するときでも計算が可能となるように (5.4) 式を修正した．step 2 では計算で積をとっているが，対数を計算し，それを最大にするような $\hat{c}$ を選択するように以下のように定義する．

$$\begin{aligned} \hat{c} &= argmax_{ci} log(P(c_i) \prod_{k=1}^{n} P(w_k|c_i)) \\ &= argmax_{ci} (logP(c_i) + \sum logP(w_k|c_i)) \end{aligned} \tag{5.9}$$

となる．本論文では和で確率を求めることによって確率が 0 になる可能性を回避し，(5.9) 式を適用した．

5.4 実験と結果

5.4.1 学習に用いるデータセット

step 1 で使用する教師データは医師の監査の元で Yahoo!JAPAN の癌カテゴリ[1]から計 31 サイトを選出し，表 5.1 の定義に従ってカテゴリ分類した．そして，分類された URL リストに対して wget プログラムを用いて個々のサイト内のウェブページを全量ダウンロードした．以降，それぞれのカテゴリの教師データの詳細に関して説明する．

- Authorized:
  Authorized の教師データとなるサイトは国立がんセンター[2]のウェブページを全量ダウンロードし，それのみを教師データとした．Authorized の教師データに国立がんセンターを用いた理由は，4 章で示したように，国立がんセンターにより発信されているがんの解説ページはがんに関する文書で標準的に使用される単語を多く含むため妥当であると考えたからである．

- Personal:
  Personal は個人が発信する闘病記や医師個人が発信するがん情報に関するウェブページが主な内容となっている．

- Media:
  Media はがん情報の書籍情報や，がん情報のポータルサイトを選出した．

- Other:
  Other はがんの漢方販売のウェブページを主に選出した．ウェブ上に存在するがんに関する商用目的のページはの多くは漢方に関するものであるため，教師データは漢方販売のページに絞った．

[1]Yaoo!カテゴリ $http://dir.yahoo.co.jp/Health/Diseases_and_Conditions/Cancer/$
[2]国立がんセンター $http://www.ncc.go.jp/jp/$

表 5.2: クローズドテストの分類実験結果

	子宮がん	胃がん	白血病	肺がん	大腸がん	average
accuracy(%)	86.4	87.0	92.6	87.5	72.7	85.2

- Noise:
  Noise はページ上の文書に “がん” という単語が出現しないものとし，計算コストを軽減させるために分類器では分類せずに，文書中に “がん” が出現しない場合は Noise とするフィルタを作成し，前処理で分類した．

最終的に得られたそれぞれのウェブページを Naive Bayesian classifier で処理するために html ファイルから html タグを外し，文書のみを抽出した．

5.4.2 解析不可能なウェブページ

これまでに説明してきたように，本章で用いる分類器はウェブページに出現する名詞に依存して分類を推定する．本研究での分類対象はウェブページであるために，文書がごくわずかで，ページ上の多くが画像データの場合がある．特にウェブページサイトのトップページの場合はページ上にあるのは，文書ではなく，そのサイトに存在するコンテンツ名のリストのみが羅列されている場合や画像のみで言語情報がまったく無いページもある．そこで，言語情報が少ないページを分析した結果，文字列総量が 150byte に満たないページに関しては本章で実装した分類器には十分な情報量ではないとみなしトレーニングデータおよびテストデータから対象外とした．したがって，本章では文字列データが 150byte 以上のウェブページを 4 つのカテゴリに自動分類することとした．

5.4.3 評価

クローズドテスト

“子宮がん”,“胃がん”,“白血病”,“肺がん”,“大腸がん”, をそれぞれ検索クエリとして Yahoo!JAPAN で検索した結果得られた上位 30 ページ (計 150 ページ) を用いてクローズドテストを行った．なお，この 30 ページはトレーニングデータに含まれているサイトである．評価尺度には全データのうちの正解したデータの割合を示す正解率 (accuracy) を用いた．クローズドテストの結果を表 5.2 に示す．

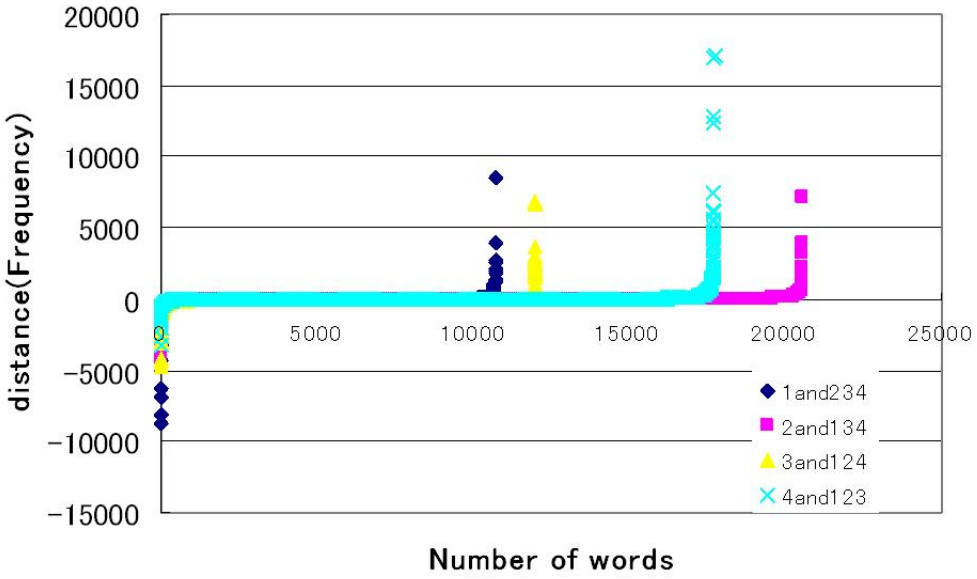

図 5.1: 各カテゴリにおける名詞頻度の比較

オープンテスト

トレーニングデータでは用いなかった疾患である “卵巣がん” を検索クエリとして Yahoo!JAPAN で検索して結果得られた上位３０サイトをテストデータとして実験した．評価尺度はクローズドテストと同様に正解率 (accuracy) を用い，83.3% の正解率を得た．クローズドデータにはやや劣るが分類器としては有用性のある精度を得られた．

5.5 WWW 上のがん情報の言語空間の考察

本章で実装した分類器のトレーニングデータとしてウェブ上のがん情報の各カテゴリにおける名詞の頻度情報を得た．この情報を分析した結果を考察する．

5.5.1 言語空間の考察

本章で実装した分類器は 5.3 節でも説明したように，ウェブページ上の名詞の頻度を素性として分類を推定する．そこで，各カテゴリ (1.Authorized, 2.Personal, 3.Media, 4.Other)

でのトレーニングデータの名詞の頻度を比較した結果を図 5.1 に示す．これは，各カテゴリにおけるそれぞれの名詞の頻度とその他のカテゴリの名詞の頻度の集合と比較したものである．例えば，1and234 であったら，2.Personal, 3.Media, 4.Other のトレーニングデータを元に，名詞の頻度をそれぞれの名詞に対して加算していき和集合を作成し，新たに 2.Personal, 3.Media, 4.Other(c-234 と呼ぶ) の 3 つを合わせた一つの集合としたカテゴリを作成する．なお，作成した和集合のそれぞれの単語の頻度は 3 で割り，平均を取ったものである．そして 1.Authorized のそれぞれの名詞の頻度から c-234 でのそれそれの単語の頻度の差を distance と呼ぶ．つまりこの値が大きくなるほどそのカテゴリで頻出する名詞であり，値が小さくなるほどそのカテゴリではあまり現れない名詞だと考えられる．

1.Authorized には “研究” という名詞が 9977 回出現する．それに対して C-234 では 1506 回出現する．この差をとると 8741 回となり，C-234 に対して 1.Authorized では “研究” が 8471 回多く出現しており，これ 1.Authorized に特徴的に現れる単語だとわかる．

逆に，1.Authorized には “漢方” という名詞が 4 回出現しているのに対して，C-234 では 8749.5 回出現している．差をとると-8790.5 となる．つまり，“漢方” が出現したらそのページは 1.Authorized ではない可能性が高いことを示唆している．ここで注目すべき点は，distance が 0 の単語が多く存在しているこである．distance が 0 ということは，つまりその名詞はウェブページの分類する際に影響していないことを意味する．

5.5.2 各カテゴリの言語的特徴

本節では 5.5.1 節で示した図 5.1 の名詞の特徴を詳しく考察する．表 5.3，表 5.4 に図 5.1 で用いたデータの distance の上位 10 名詞と下位 10 名詞を示したものである．1and234 の表の考察を述べる．特徴的な名詞は distance が一番大きい “研究” と，逆に一番小さい “漢方” である．これは個人や企業が発信する情報の質の違いを表している．これは本研究の目的でも述べたように，がん患者は情報の選択に困難を強いる原因となると考えられる．1.Authorized で化学療法などの専門的な名詞がよく使われるのに対し，4.Other では漢方の説明が多いことが予想される．がんは治療法が確立されていない疾患であるため，様々な治療法がウェブ上で説明されるのは当然のことであるが，この問題は命に関わる問題なので深刻である．

また，1.Authorized には “相談” という名詞がほとんど出現してないことが示されている．この名詞は主に 4.Other で頻出している単語である．1.Authorized では，各がんの解説や症状をまとめて解説しているページが多いが，がん患者にとって専門的な文書は難

解である．また，病気の進行や，段階によって患者の悩みや知りたいことは病期や病状によって様々なことがある．今日ウェブ上でがん患者に求められているのは，がんに関する情報に加え，気兼ねなく相談できるようなシステムが必要されている可能性が高いことが示唆された．今後の研究として，1.Authorized の情報だけでは足りないような付加的な情報を 2.Personal の体験談や医師個人の発信する情報と組み合わせて情報を提供するシステムを研究していきたいと考えている．

人称代名詞の使い方にも違いが現れた．例えば，“私” という名詞は一人称で用いられる単語であり，1.Authorized で使われることは少ない．“私” は闘病記や体験記に特徴的に使われる名詞である．そのほかに，“先生” という名詞も一般的には患者が使う名詞であり，1.Authorized では “医師” という名詞を用いる．これは一例に過ぎないのだが，医師が記述するがんのウェブページと個人が記述するそれでは同じ内容を述べていても使用する単語に違いがあることを意味している．

表 5.3: 各カテゴリにおける特徴的な単語 (1and234,2and134)

1and234			
名詞	distance	名詞	distance
研究	8471	漢方	-8790.5
一覧	3906	相談	-8152.5
国立	2782.75	子宮	-6928
がんセンター	2779.75	シート	-6219.75
更新	2552.5	私	-4354.75
遺伝子	2051.75	抗がん剤	-3415
先頭	2020.75	治療	-3380.5
目次	1914.75	体	-3223.25
問い合わせ	1764.75	薬局	-3132.25
内容	1278.5	医学	-3000
化学療法	1244	卵巣	-2863.25

2and134			
名詞	distance	名詞	distance
私	7216.25	研究	-4987.75
入院	3917.75	相談	-4558.75
病院	3905.75	漢方	-3888.75
検査	3240	シート	-3066
自分	3214.25	情報	-2086.5
先生	2336.25	一覧	-2069.75
海外	1875	抗がん剤	-2062.5
手術	1871.75	内容	-2034.75
これ	1816.5	必須	-1739.5
人	1805.75	薬局	-1599

表 5.4: 各カテゴリにおける特徴的な単語 (3and124,4and123)

3and124			
名詞	distance	名詞	distance
必須	6875.5	研究	-4725.25
記入	6763.5	相談	-4473.75
番組	3656	漢方	-4253.75
情報	2971	子宮	-4155.5
家族	2778.5	シート	-3124.75
本人	2707	冬虫夏草	-2953
患者	2672.25	治療	-2462.25
全角	2570.5	細胞	-2235
個人	2461.75	抗がん剤	-2152.5
ホームページ	2393	一覧	-2072.25

4and123			
名詞	distance	名詞	distance
漢方	17095	研究	-3340
相談	16965	病院	-2605.5
子宮	12812	国立	-1970.75
シート	12417.75	一覧	-1776
抗がん剤	7430	必須	-1607
薬局	6314.75	記入	-1565.25
体	6096.25	医療	-1407.75
治療	5594	更新	-1383.5
医学	5463	がんセンター	-1361.75
卵巣	5007.25	全角	-1310

第6章　ウェブ形態を用いたがん情報の分類

6.1　分類にウェブ形態情報を用いる目的

5 章で示したように，がん情報の自動分類はウェブページの文書中に出現するすべての名詞の出現頻度を用いて分類することで 8 割近い分類精度が得られた．しかし，図 6.1 に示すように Other(商用情報など) 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．図 6.1 から Other のページが少ない “白血病” は分類精度が良く，Other のページ数が増えるほど，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．

- Other 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箇所が混在しているページがあるため．具体的な例を図 6.2 に示す．
- 個人や業者ががんの疾患を解説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を参照して記述したウェブページがあるため．具体的な例を図 6.3 に示す．

図 6.2 の場合，名詞の出現頻度を用いて分類すると，疾患の解説部分のに強く作用されてしまい Other であるページが，Authorized や Personal のページであると誤判別してしまう可能性がある．また，図 6.3 の場合は，文書の引用や参照をしているため，似通った名詞の生起頻度から分類器は誤判別してしまう可能性がある [17].

従って，本章では言語情報だけを素性として分類し，誤判別することを避けるために，言語以外にウェブページの分類に有効な素性を発見し，その有用性を検討する．

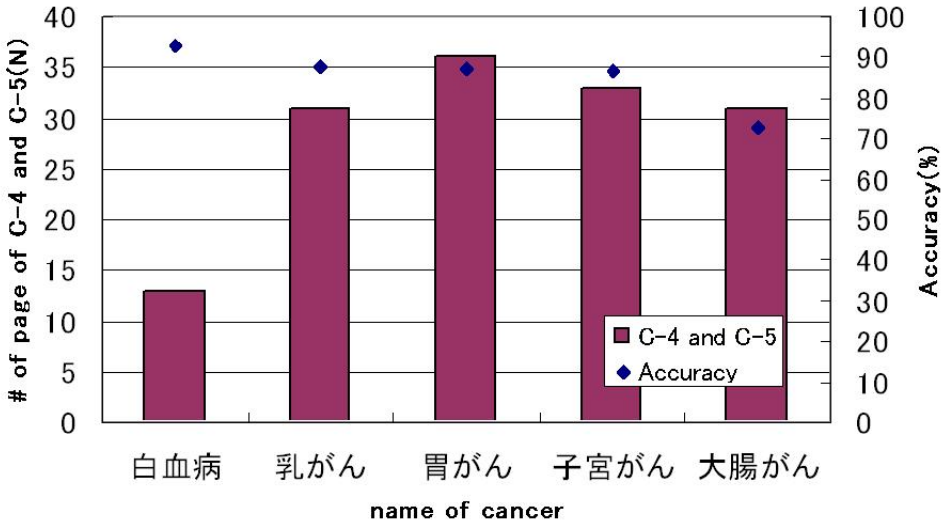

図 6.1: Other のページ数と分類精度の関係

6.2 提案手法

6.2.1 基本的なアイデア

がん情報では，CII のカテゴリ間で情報の質が違うため，言語以外にも特徴が現れることが推測される．数多くのがん情報のウェブページを閲覧する中で，がん情報は各カテゴリ間で言語以外にもページを見た瞬間の視覚的な特徴があることに気がついた．

例えば，Authorized のページでは疾患を詳しく解説するために jpeg イメージを使う頻度が高くなる可能性が高い．Personal のページでは frame タグが使用されて複数のページからウェブページが構成されているものや，midi などを用いたオーディオファイルをコンテンツに含めいてること．Other では広告を目的としたページが多いため，ウェブページを構成する html ファイルの総容量が大きくなることや，販売目的であるページは販売するためのプログラムを JavaScript で設置しているページが多く見られることなどである．具体的な例を図 6.4 に示す．

しかし，これだけの特徴量だけでは，分類は困難であろうことは予測できる．そこで本章ではウェブページには表面的に現れないが，ウェブページの内容を表す head 要素に

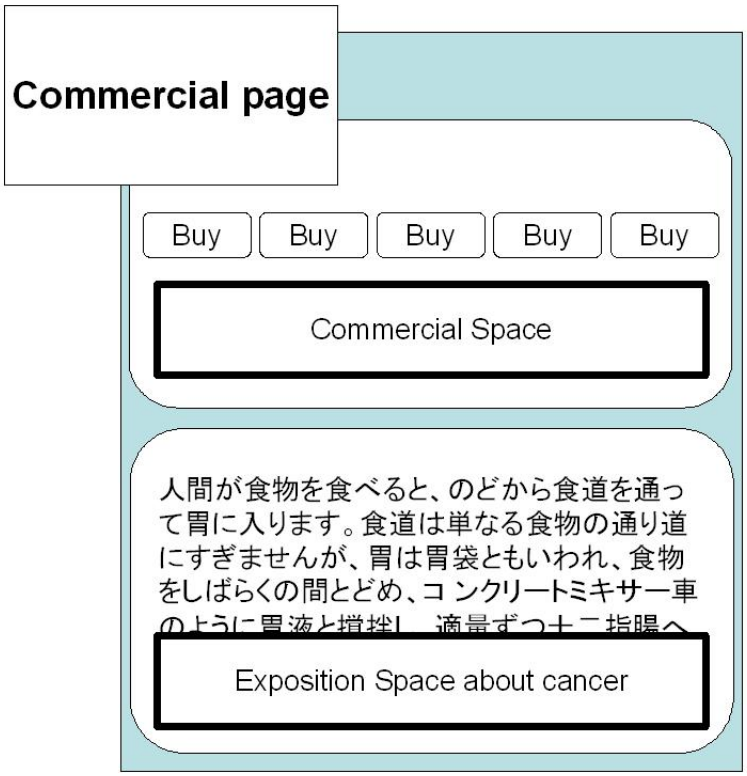

図 6.2: Other(Commercial) のウェブページの例

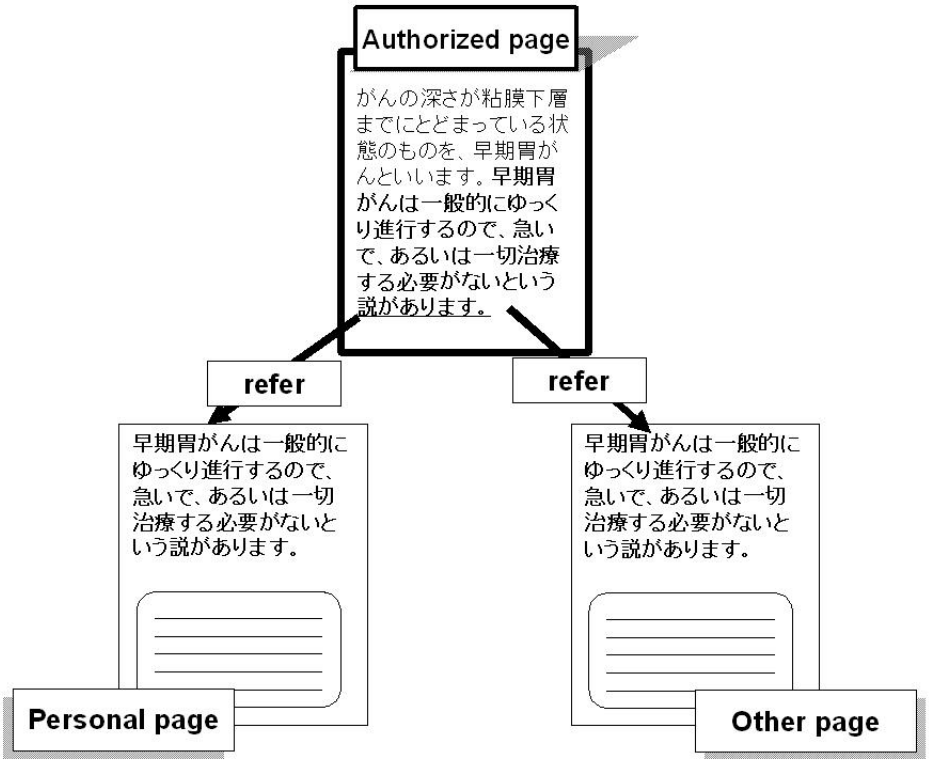

図 6.3: Authorized のページを参照している例

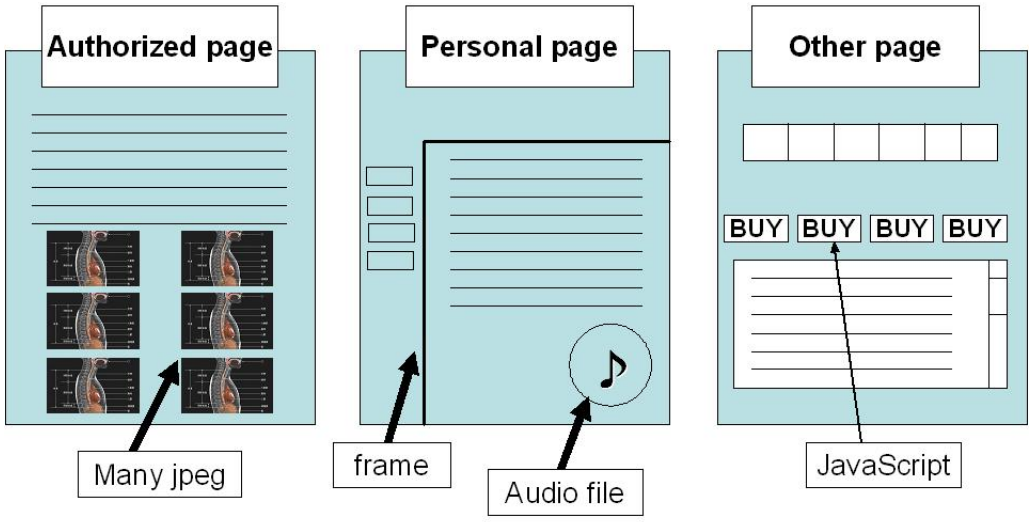

図 6.4: 基本的なアイデアの具体的な例

着目した．head 要素にはウェブページの title やウェブページのキーワード，要約などが記述される．head 要素の多くはウェブクローラーに効率的にクローリングされるためにウェブページの作成者が記述する．これらの情報は直接的には人間の視覚に認知されないが，キーワードや要約などの情報はページの内容を要約された情報であり，ウェブページを認識するために特徴量が大きいことが推測される．以降本章で用いる各素性を説明し，統計的手法を用いてウェブの形態的な素性の有用性を検討する．

6.2.2 分類に用いる素性

以上の検討から本章では，提供されているコンテンツの形態素解析を精密化しても分類不能である悪意を持ったコンテンツの検出に役立つ可能性のある，コンテンツ特徴量 (特に URL に含まれる客観的計測項目) をウェブページの評価指標として与えることを目的とする．ウェブページ上の文書中に出現する言語に関する素性として専門用語比，ならびに URL ツリーを全量ダウンロードして客観的に計測可能なウェブの形態に関する素性 (コンテンツ量などのデータ構成に関する各種客観的計測項目およびヘッダから客観的に

設定可能な情報) をできるだけ広範囲に (20 項目に関して) 検討し，実用上有用なパラメータを検討することとした．本章で検討する 20 項目を表 6.1 に示した．以降，この 20 値の素性に関して詳しく説明していく．

言語に関する素性

- 専門用語比
  専門用語比 ($techniq\_rate$) 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 Chasen + ipadic を使用した．なお，専門用語が認識できるように，ipadic には中川が作成したがん専門用語集 3316 語と医学専門用語約 59533 語 [21] を追加した．専門用語比の式を示す．$f(T_j)$ 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．$f(W_k)$ 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．

$$techniq\_rate_i = \frac{\sum_{j=1} f(T_j)}{\sum_{k=1} f(W_k)} \tag{6.1}$$

ウェブ形態に関する素性

ウェブ形態とはウェブページを構成する html ファイルの総容量やイメージファイルの総数などといったウェブページを構成する要素を計測し，数値的にあらわしたものである．本研究で素性として用いるウェブ形態を構成情報，haed 要素情報，その他の付加情報にわけて説明する．

- 構成情報の素性

  1. html ファイル総量 ($html\_size$)
  2. html ファイル総数 ($html\_number$)
  3. jpeg 総量 (jpg_size)
  4. jpeg 総数 (jpg_number)
  5. gif 総量 (gif_size)
  6. gif 総数 (gif_number)
  7. png 総量 (png_size)

表 6.1: 実験に用いるウェブ形態素性 20 値

素性名	説明
専門用語比 ($techniq\_rate$)	文書中に生気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.
html ファイル総量 ($html\_number$)	ページを構成する全ての html ファイルの総容量 (byte).
html ファイル総数 ($html\_size$)	ページを構成する全ての html ファイルの総数.
jpeg 総容量 (jpg_size)	ページ上にある jpeg イメージの総量 (byte).
jpeg 総数 (jpg_number)	ページ上にある jpeg イメージの総数.
gif 総容量 (gif_size)	ページ上にある gif イメージの総量 (byte).
gif 総数 (gif_number)	ページ上にある gif イメージの総数.
png 総容量 (png_size)	ページ上にある png イメージの総容量 (byte).
png 総数 (png_number)	ページ上にある png イメージの総数.
title 文字数 ($title\_size$)	ページの title 要素の文字数.
author 文字数 ($author\_size$)	author 要素の文字数. author はページの作成者を記述する.
description 文字数 ($description\_size$)	description 要素の文字数. description はページの要約を記述する.
keywords 総数 ($keywords\_size$)	keywords 要素の内にあるキーワードの総数.
head 要素数 ($head\_elements$)	head 要素内にある子要素の総数.
JavaScript	ページ上で javascript が使用されているか.
CSS	ページ上で CSS(スタイルシート) が使用されているか.
flash	ページ上で flash が使用されているか.
audio	ページ上で audio ファイルがあるか.
depth	ドメインネームからの深さを計測したもの.
ドメイン情報 (top_domain)	ページのトップドメイン. 具体的には co.jp や ac.jp など.

8. png 総数 (*png_number*)

- head 要素の素性

  head 要素とはウェブページのヘッダをあらわすものである [10]．head 要素には title 要素を子要素として必ず含む．その他に，文書の無いように関する meta 要素などがある．本研究で素性として取り入れた head 要素の素性を説明する．

  1. title 文字数 (*title_size*)
  2. author 文字数 (*author_size*)
     author と meta タグの一要素であり，ウェブページの作成者や所属や所属などを記述するためのタグである．
  3. description 文字数 (*description_size*
     description は meta タグの一要素であり，ウェブページの内容の要約を記述するためのタグである．
  4. keywords 総数 (*keyword_size*) keywords は meta タグの一要素であり，ウェブページの内容に関するキーワードを記述するためのタグである．
  5. head 要素数 (*head_elements*) これは head 要素にある子要素数である．head 要素の中には作成者によって子要素を任意の数を記述することができる．

- その他の付加情報の素性

  1. JavaScript が使用されているか (*javascript*)
  2. CSS(スタイルシート) を使用しているか (*css*)
  3. flash を使用しているか (*flash*)
     flash とは Macromedia 社が開発した，音声やベクターグラフィックスのアニメーションを組み合わせてウェブコンテンツを作成するソフトによって作成されたコンテンツのことである．
  4. audio ファイルが使用されているか (*audio*)
     ホームページに使用されるオーディオファイルの多くは midi(Musical Instruments Digital Interface) と呼ばれる，楽曲データをやりとりするための規格が用いられる．
  5. ファイルの深さ (*depth*)
     分類対象のウェブページのドメインネームからの深さを計測したものである．

例えば，ドメインネームの直下におかれている index.html であれば，深さ 1 とする．

6. ドメイン情報 (*top_domain*)
ドメイン情報は分類対象のウェブページのトップレベルドメインのことである．具体的には “co.jp” や “ac.jp” などのことである．一般的には組織によって使用できるトップレベルドメインが異なる．

6.3 統計を用いたウェブ形態の有用性の検証

現在知られている分類アルゴリズムは，ベクトル化するときに用いる変数の統計学的特徴により，分類精度が変動することが知られている．特に問題となるのは，分類器の用いるアルゴリズムに適切なベクトル化変数を選択しなければ，分類精度が低下する場合がある．そこで，前項で列挙した素性のうち，カテゴリ名を従属変量として一般線形モデル (GLM) を適用してこれら諸値から素性選択を行い，分類精度を高めるものを検索することとした．

6.3.1 データセットの固定と教師データの作成

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 URL 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 1000 として URL リストが提供されているが今回はその中で，上位 100 ページ (計 700) を対象とした．それぞれの URL に従い wget を用いて対象とする URL ツリーデータを全量ダウンロードした．ページが存在しないものなどを除外し，計 675 ページを実験に用いるデータセットとして固定した．

本データを対象として，医師の資格を持つ者により，定義したカテゴリ（1: Authorized, 2: Personal　および 3: Other）に従って CL-Score を作成した．各疾患でのスコアの分布とページ数を表 6.2 に示した．

6.3.2 ウェブ形態素性諸値の検討

675 ページを対象として，前項で述べた，専門用語数比 (*techniq_rate*)，ウェブ形態素性諸値 8 値（$html\_number, html\_size, jpg\_size, jpg\_number, gif\_size, gif\_number, png\_size,$

表 6.2: 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

png_number)，Header 素性諸値 5 値（$title\_size, author\_size, description\_size, keyword\_size, head_elements$），ならびにその他の素性情報 4 値（$javascript, css, flash, audio, depth$）の合計 19 変数について検討した．ドメイン情報諸値（16 値）は，一元的数値化が困難であることから，GLM の説明変数にはせず，分類器に直接”ac.jp”, ”co.jp”等の文字列として入力することとした (図 6.6)．念のため，ドメイン情報諸値単独の C4.5, SVM での CL-Score の分類精度を求めたが，F-Measure で 50% 程度であったため，本変量に統計的に選択された変量を加えて分類した場合の目標は精度 50%以上とした．

表 6.3 に 675 ページにおける，計測項目 15 変数の平均値と標準偏差を示す．これら変数の分布は従来知見で報告されているように指数分布を示した．変数のうち，連続的ではなく尺度的変量である $javascript, css, flash$ ならびに $audio$ の 4 変数について CL-Score との関係を χ 二乗検定で検討した．結果を表 6.4 に示した．その結果，変数”$flash$”を除く 3 変数が CL-Score と有意であった．

表 6.5 に，CL-Score を従属変数とし，これら 19 変数を説明変数とし，Stepwise($p < 0.05$ で選択，$p > 0.1$ で除外) 法を用いて作成したモデルの要約を示した．モデル f の重相関係数 R は 0.44 であり，$technique\_rate, description\_size, depth, author\_size, jpg\_size, javascript$ が選択された．予測値と CL-Score の関係を図 6.5 に示した．以上の検討からこれら 6 値とドメイン情報諸値（図 6.6）を分類のためのベクトル化変数として用いることとした．

表 6.3: 各素性の平均値と標準偏差

N of Cases	675	
	Mean	Std.Deviation
jpg_size(K Byte)	19.99	52.60
jpg_number	1.49	3.77
gif_size(K Byte)	28.74	66.01
gif_num	10.36	12.91
png_size(K Byte)	0.44	7.15
png_number	0.11	1.19
$html\_size$(K Byte)	23.91	22.76
$html\_number$	1.20	0.70
$techniq\_rate$	0.04	0.03
$description\_size$	50.98	102.76
$author\_size$	2.44	14.42
$title\_size$	34.96	23.24
$keyword_size$	6.42	18.75
$head\_elements$	4.46	3.19
$depth$	2.51	1.42

表 6.4: 4 変数に対する χ 二乗検定の結果

	Value	df	p-Value
javascript	11.99	2	0.00
css	17.98	2	0.00
flash	5.1454	4	0.27
audio	7.49	2	0.02

表 6.5: 18 変数に対し Stepwise 法を用いて変数選択をした結果

Model	R	R Square
a	0.33	0.09
b	0.36	0.13
c	0.40	0.16
d	0.41	0.17
e	0.43	0.19
f	0.44	0.20

a: Predictors:(Constant), VAR12
b: Predictors:(Constant), VAR12, VAR13
c: Predictors:(Constant), VAR12, VAR13, VAR18
d: Predictors:(Constant), VAR12, VAR13, VAR18, VAR14
e: Predictors:(Constant), VAR12, VAR13, VAR18, VAR14, VAR4
f: Predictors:(Constant), VAR12, VAR13, VAR18, VAR14, VAR4, VAR19
VAR12=techniq_rate, VAR13=description_size, VAR18=depth
VAR14=author_size, VAR4=jpg_size, VAR19=javascript

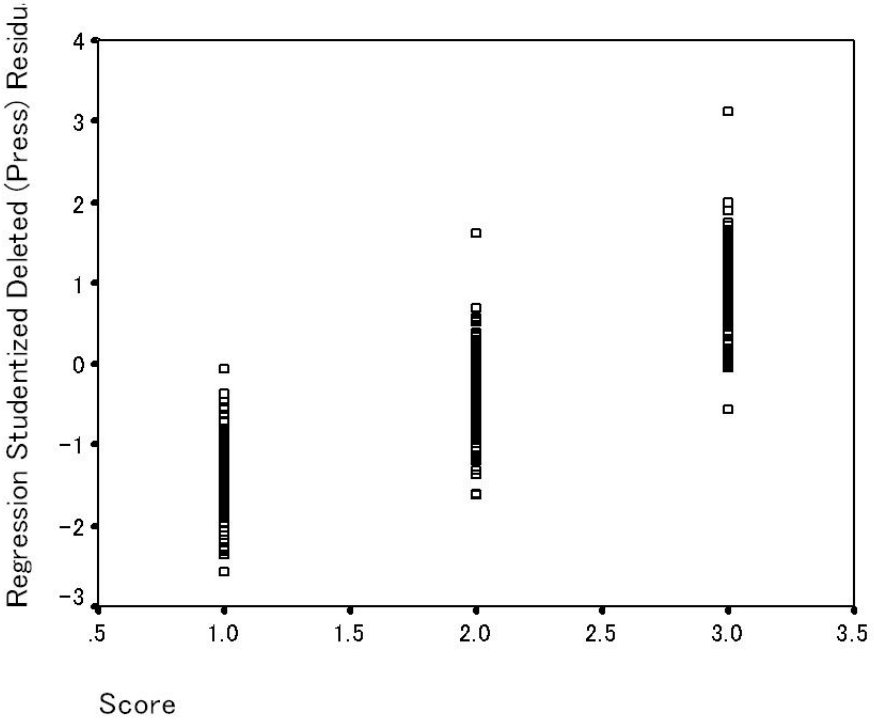

図 6.5: GLM で選択された変数 6 値の予測値の Scatter plot

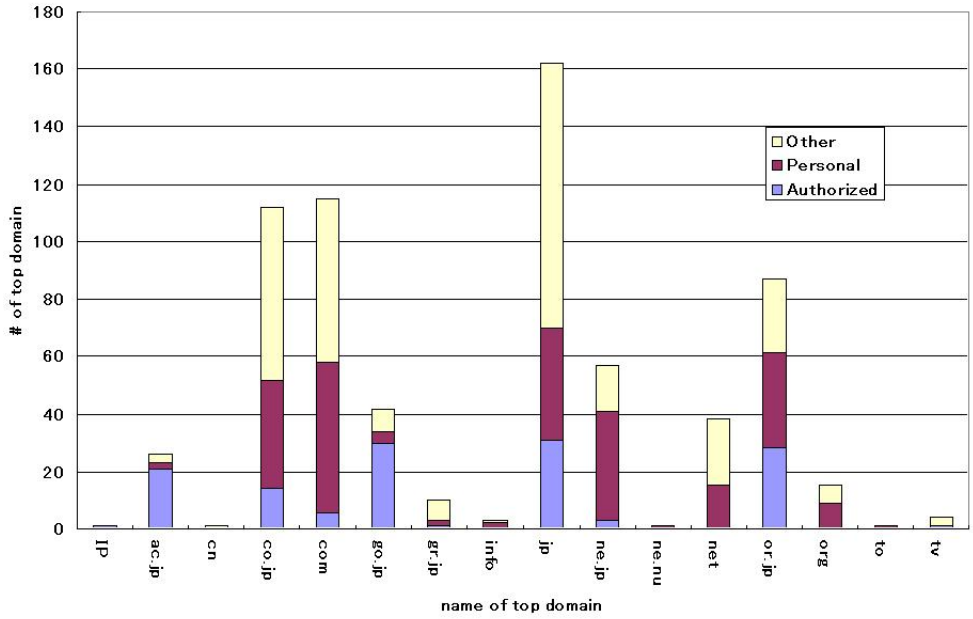

図 6.6: 各カテゴリにおける top domain の頻度

6.4 評価実験

6.4.1 分類カテゴリの定義

カテゴリの定義には，CII を修正し，以下の 3 つのカテゴリを用いることとした．

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．

6.4.2 実験に用いた 2 種類の素性セット

本章で提案した素性の有用性を検討するために，以下の 2 種類の素性セットを作成し，分類器にそれぞれの素性セットを与えて分類した結果の分類精度を考察する．

- **重回帰分析によって選択されたウェブ形態素性 6 値とドメイン情報 (計 7 値の素性)**
  重回帰分析を行い，一般線形モデルから選択された計 6 値の素性 ($jpg\_size, techni1\_rate, description\_size, author\_size, javascript$) と top_domain を素性セットとした．

- **ウェブ形態情報全て (計 20 値の素性)**
  本研究で提案した計 20 値の素性全てを素性セットとした．

6.4.3 評価

6.3.1 節で説明した計 675 ページのデータセットと， 6.4.2 節で説明した 2 種類の素性セットを用いて評価実験を行った．比較検討をするために，分類器は SVM [8] と C4.5 [6] の 2 つの分類器を使用し，10 交差検定法を用いた．SVM，C4.5 ともに weka [9] により実装された分類器を用いた．

表 6.6: 各素性セットの素性の数

	number of attributes
GLM selected attributes	7
All features	20

表 6.7: 各素性セットで分類した結果

	C4.5	SVM
GLM selected attriubutes	0.64	0.55
All features	0.62	0.6

まず各素性セットの素性数を表 6.6 に示す．提案した素性はそれぞれ，7 個，20 個であり，とても小さな素性セットで分類する．それぞれの分類器の分類結果を表 6.7 に示す．この表の値はそれぞれカテテゴリでの F-measure の平均値を示したものである．F-measure の式を式 6.2 に示す．この結果からわかるように，SVM の場合は提案した 20 値すべての素性を分類に用いたほうが良い分類精度を得た．C4.5 においては，とても少ない素性で，SVM，C4.5 ともに 6 割以上の F-measure を得ることができた．特に C4.5 に関しては，重回帰分析で素性の数を約 1/3 に減らしたにも関わらず，より良い分類精度を得られたことから, この 7 個の素性が分類に有効であることが示唆された．

$$F - measure = \frac{2PR}{P+R} \tag{6.2}$$

$$P = Precision, R = Recall$$

6.4.4 実験結果の考察

6.4.3 の実験でウェブの形態情報を用いた 2 種類の素性セットを用いて分類した結果を示した．実験結果からウェブ形態を用いた分類は名詞の頻度を用いた分類よりもやや精度は劣るが，有用性があることが示唆された．オープンド メインでのウェブページの分類問題で本手法が効果的に働くかなど，今後もウェブ形態情報とウェブページの分類に与える効果を検討するべき課題がある．また，言語情報とウェブ形態情報を組み合わせた分類モデルでより良い効果が得られるのではないかと期待している．

第7章　ウェブ形態情報と言語情報を用いたがん情報の分類

本章では，6章で示した，客観的に計測可能な素性であるウェブの形態的な素性7値と，言語の計量的特徴に関する素性6値を従来文書分類で行われてきた形態素解析を行い文書中に出現する名詞の頻度に加えて分類した結果，分類精度に与える影響に関して考察する [18].

7.1　提案手法

本研究では6章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) と言語の計量的特徴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の計量的特徴に関する素性について説明する．

7.2　ウェブ形態に関する素性

本研究では6章で示された以下の7値の素性をウェブ形態として用いる．

- *jpeg* 総量
- *author* 文字数
- *description* 文字数
- Javascript の有無
- ファイルの深さ
- ドメイン情報

7.3 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．

7.3.1 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 の重みをと考える．つまり、個々の名詞の重みを w_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した [14].

7.3.2 言語の計量的特徴に関する素性

本研究では，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 [13]．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 7.1 に示す．以下，それぞれ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 7.2 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.

表 7.1: 実験に用いてる言語形態素性 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.

表 7.2: 言語形態素性の平均値と標準偏差

number 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	142.20
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

7.4 評価実験

7.4.1 カテゴリの定義

分類に用いるカテゴリの定義は 6.4.1 節で定義した 3 つのカテゴリを用いることにした (1.Authorized 2.Personal 3.Other).

7.4.2 実験に用いたデータセット

実験に用いるデータセットも 6 章と比較検討するために同様のものを用いることにした．よってデータセットの詳細は 6.3.1 節を参照されたい.

7.4.3 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 [9] による SVM 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．

7.4.4 実験結果

表 7.3 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 ，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 7.4 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表 7.5 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．

表 7.3: 分類実験の結果

一般名詞			
Category	Precision	Recall	F-Measure
Authorized	0.72	0.61	0.66
Personal	0.66	0.67	0.67
Other	0.75	0.79	0.77

一般名詞 + ウェブ形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.75	0.68	0.71
Personal	0.68	0.69	0.69
Other	0.77	0.80	0.79

一般名詞 + 言語形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.71	0.60	0.65
Personal	0.66	0.67	0.66
Other	0.75	0.80	0.77

一般名詞 + ウェブ形態素性 + 言語形態素性			
Category	Precision	Recall	F-Measure
Authorized	0.74	0.66	0.7
Personal	0.68	0.69	0.68
Other	0.77	0.80	0.78

表 7.4: 分類の結果 (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	0.73
一般名詞 + 言語形態素性	0.69
一般名詞 + ウェブ形態素性 + 言語形態素性	0.72

表 7.5: 各素性セットの素性数と学習モデルを作成するのに要した時間

素性セット	number of feature	Time(modesl)
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

第8章　おわりに

8.1　まとめ

本研究では，一般的に使用される検索エンジンでは無秩序に出力されるがんに関する情報に対し，がんに関する専門知識がない一般人にもがんの情報を正しく選別できるようにするために次の検討を行った．

- 言語情報を用いたがん情報の分類
  Naive Bayesian classifier を実装し，がんに関するウェブページの文書中に出現する名詞の頻度を素性として分類実験を行った．分類実験の結果クローズドテストで約85%，オープンテストで約83%の Accuracy を得たことから，がん情報の文書中に出現する名詞の頻度が分類に有効であることが示された．また，がん情報の言語空間の考察から分類に寄与している名詞が限られていることが示されたことから，今後分類にもっとも有効な素性選択の手法に関して検討していきたいと考えている．そして，それぞれのカテゴリにおいて中心的に語られている概念に違いがあることも示された．例えば，Authorized のページでは “研究”，“化学療法” という単語がよく使われていることから，医学的根拠を持った治療法の解説をしているページが多いことが示唆され，Other では “漢方” の頻度がとても多く，このような情報の内容の違いが患者の情報取得を困惑させてしまっている可能性があることを示唆した．

- ウェブ形態情報を用いたがん情報の分類
  言語情報を用いたがん情報の分類で，Other のページが多いほど分類精度が低下していることが示された．この原因は商用誘導を企むページががんの疾患を解説している箇所と，販売を目的とした箇所を一つのページに混在しているケースがあることや，がんの疾患を解説する際に Authorized のページを引用するケースがあることだと考えた．そこで，言語情報以外にウェブページに特有に現れる素性20値を統計的手法で有用性を検証した結果，専門用語比，description 文字数，ファイルの深さ，author 文字数，jpeg 総容量，javascript が使用されているかどうか，ドメイン情報

の 7 値が有用であることが示された．7 値のウェブ形態素性を用いてがん情報を分類したところ，分類器 C4.5 で 0.64%の F-Measure を得た．言語素性に比べとても少ない素性で分類したにも関わらず，6 割以上の分類精度を得られたことから，ウェブの形態的な素性もがん情報においては分類に有効であることを示した．

- ウェブ形態情報と言語情報を用いたがん情報の分類
  選択されたウェブ形態素性 7 値に言語情報を組み合わせた自動分類を検討した．言語に関する素性は名詞の頻度に加え，文書中の句読点の数などの言語の計量的特徴に関する素性 6 値を検討した．それぞれを組み合わせた 4 種類の素性セットを用いて分類実験を行った結果，一般名詞とウェブ形態素性を組み合わせた素性セットでの分類がもっとも良い精度を得た．このことからウェブ上のがん情報の分類は，文書中の言語情報に加え，ウェブの形態的な情報を素性として用いる手法が本研究においてはもっとも有効であることが示された．本研究で用いた言語の計量的特徴は，分類精度を低下させてしまう結果となった．

8.2 今後の研究と課題

- 素性選択の検討
  本研究では一般名詞 + ウェブ形態素性の組み合わせでの素性セットがもっと良い分類精度を得た．一般名詞に対する有効な素性選択に関しては今現在も検討中であるため，今後一般名詞に関してウェブ上のがん情報の分類に最も有効な素性選択の手法の検討及び，選択された素性の考察を行わなければならない．

- より大規模なウェブページ群に対しての分類実験
  本研究用いたデータセットは 7 疾患に関するウェブページ 675 ページを用いた．しかし，今後実際に提供するシステムを想定した研究をするためにがんに関する全ての疾患 (計 59 疾患) を網羅した分類実験を行わなければならない．現在，59 疾患を検索クエリとして得られた 44477URL のがんに関するウェブページを得ている．今後は 44777URL のがん情報に対して教師データの作成し，分類手法に関して検討していく予定である．

謝辞

本研究を進めるにあたり，島津明教授，白井清昭助教授, 鳥澤健太郎助教授，中川晋一助教授には，数多くの御教示を頂きました．また，本研究に関して，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，中村誠助手, 山田寛康元助手ならびに島津・白井研究室の皆様方には，研究に関する貴重な支援をして頂きましたことを心より感謝致します．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．

参考文献

[1] C.Cortes and Vladimir N.Vapnik,
Support Vector Networks, Machine Learning,Vol.20,pp.273-297, 1995.

[2] Friedman.N,Geiger.D,Goldszmidt.M, Bayesian network classifiers, Machine Learning, 29(2-3), 131 – 163, 1997.

[3] Gray Malet, Felix Muonz, Richard Appleyard, William Hersh, A Model for Enhancing Internet Medical Document Rtrieval with "Medical Core Metadata", Journal of the American Medical Informatics Association, Volume 6 Number 2, 1999.

[4] Hiroshi Nakagawa, Automatic Term Recognition based on Statistics of Compound Nouns Terminology, Vol.6, No.2, pp.195 - 210, 2000.

[5] NHK, NHK SPECIAL HOME APGE,
http://www.nhk.or.jp/special/libraly/06/10001/10107.html

[6] J.Ross Quinlan, C4.5: programs for machine learning, Morgan Kaufmann, 1944.

[7] Mehran, Sahami, Susan Dumais, David Heckerman, Eric Horvitz, A Bayesian Approach to Filtering Junk E-mail, AAAI'98 Workshop on Learning for Text Categorization, July 1998.

[8] Susan Dumais, Hao Chen, Hierarchical claddification of Web content, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retieval(SIGIR2000), pp.256 – 263, Athenes, Greece, July 2000.

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

[10] W3C, W3C Technical Reports and Publications,
http://www.w3m.org/TR/.

[11] 阿部倫子，田中久美子，中川裕治, コメントを用いた映画の分類, 情報処理学会自然言語処理研究会，NL-150,pp.105 – 110, 2002.

[12] 落谷亮，WWW ページの分類におけるテキストの特徴分析手法，情報処理学会研究報告自然言語処理 118 – 14, pp.85 – 90, 1997.

[13] 金明哲，村上征勝，永田昌明，大津起夫，山西健司, 統計科学のフロンティア 10 言語と心理の統計, pp3 - pp57．岩波書店,2003.

[14] 北研二， 言語と計算 – 4 確率的言語モデル， 東京大学出版会,1999.

[15] 木村俊也，中川晋一，三角真，島津明，山岡克式，酒井善則，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 - Bayesian classifier を用いたがん情報コンテンツの分類 -, 電子情報通信学会第 17 回データ工学ワークショップ/第 4 回日本データベース学会年次大会 (DEWS2006),2006.

[16] 木村俊也，中川晋一，三角真，山岡克式，酒井善則，島津明， Web 上の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 (NLP2006),2006.

[17] 木村俊也，中川晋一，三角真，島津明，山岡克式，酒井善則， ウェブの形態情報を用いたがん情報の分類, 電子情報通信学会第 18 回データ工学ワークショップ/第 5 回日本データベース学会年次大会 (DEWS2007),2007.

[18] 木村俊也，中川晋一，三角真，山岡克式，酒井善則，島津明，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，言語処理学会第 12 回年次大会 (NLP2006),2006.

[19] 中川晋一, 木村俊也，三角真，島津明，山岡克式，酒井善則, 介入的手法によるがん情報取得適正化に関する検討，電子情報通信学会第 17 回データ工学ワークショップ/第 4 回日本データベース学会年次大会 (DEWS2006), 1b-i10．2006.

[20] 中川晋一, 木村俊也，三角真，島津明，山岡克式，酒井善則, 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 Vol.5 No.1, pp 21 – 24, 2006.

[21] 中川晋一，木村俊也，三角真，島津明，山岡克式，酒井善則， Web がん情報評価のための単語集合の作成，電子情報通信学会第 18 回データ工学ワークショップ/第 5 回日本データベース学会年次大会 (DEWS2007),2007.

[22] 中川晋一，木村俊也，三角真，島津明，山岡克式，酒井善則, ウェブがん情報空間推定のための単語集合に関する検討, 言語処理学会第 13 回年次大会 (NLP2007), PA2-3, 2007.

[23] 長沼潔, 速水悟，医療分野における Web 文書からの話題抽出方法，THe 19th Annual Conference of the Japanese Society for Artificial Intelligence,2005.

[24] 松本裕治，北内啓，平野善隆，松田寛, 形態素解析システム「茶筅」version 2.3.3 使用説明書，奈良先端科学技術大学院大学松本研究室, 2003 年 8 月.