ウェブの形態情報を用いたがん情報の分類

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

†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 〒152-8550 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

あらまし 近年増加が著しいウェブ上のがんに関する情報は，客観的な評価がされていない膨大な情報から必要な情報を選択することが困難であるとされている．この問題を解決するために我々は，がん情報を１．Authorized(専門医が記述した情報)，２．Personal(個人が記述した情報)，３．Other(販売を目的とする情報)の 3 種類のクラスに自動分類して提供する手法を検討してきた．一般的にウェブページは文書中に出現する単語頻度や共起頻度などを素性として用いるのが有効である．しかし，がん情報では商用誘導を企てたページや，専門医が記述する文章と酷似したページが存在するために言語モデルのみを素性として分類すると分類を誤る場合がある．本稿ではページ上に存在するイメージの数やページの総量といったウェブページの形態情報(ウェブ形態情報と呼ぶ)を分析し分類に有効な素性を発見する．そしてウェブ形態を用いて分類した結果，ウェブ形態が分類精度に与える効果を検討する．

キーワード ウェブマイニング，情報検索，文書分類，素性選択

1. はじめに

がんに関する情報は，未だ標準的治療が確立されていないことから，がん患者や家族にとって，ウェブで提供されている情報は，新規性，経済性，提供されている範囲，検索の利便性などの点で重要である．しかしながら，これら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事が難しく，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[1].医学分野でのウェブマイニングを行い病理の症状や原因を抽出する研究[2]が報告されているが，がんに特化した報告は未だない．また問題の解決のためにウェブページでの情報発信に対して倫理基準を適応しようとする例[3]もあるが,処理すべき情報量が多いこと，判定プロセスの透明性を確保するために機械化すると解析されるという悪循環がある．通常の検索エンジンで得られる情報はさまざまな質のものを並列に提示することが問題であり，提供された URL リストを機械的に分類し，評価指標を与えることを検討してきた[4]-[7].

2. システム概要

本研究ではがん患者がウェブからがんに関する情報を検索する際に情報選択の手助けをするため，がんに関するウェブページを書き手による分類を行う．なお,本論文ではがんに関するウェブページをがん情報と呼ぶ．目指すシステムの全体像を Fig.1 に示す．本システムはあらかじめがん情報を WWW から収集しておき，それぞれのウェブページの書き手を推定し同定する．検索者はシステムに対して検索クエリを与え，書き手が同定されたウェブページの集合から検索クエリに一致するウェブページを出力する．

2.1. 分類カテゴリの定義

がん情報の分類は，中川ら[4]による CII(Cancer Information Index)を修正し，以下の３つのカテゴリを用いる．

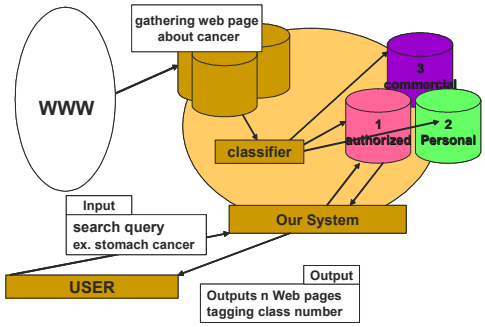

Fig.1 Overview of system

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可能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がん情報．この情報は信頼性が低いものとして提供する．

3. 本研究の目的

従来文書の自動分類は文書中に出現する名詞の出現頻度を用いて分類することで良い分類精度を得ている．がん情報もこれと同じよう言語情報を用いて分類

することは不可能ではない [5].

しかし，がん情報では特に悪意を持って商用誘導を企てたウェブページが存在するため，言語情報だけでは誤分類してしまう可能性がある．商用誘導を企むページでは，ウェブページ上に販売を目的としたスペースと，がんの疾患を解説するためのスペースが混在しているケースが見受けられる(Fig.2)．このような場合，名詞の出現頻度を用いて分類すると，疾患の解説部分の頻度情報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，Personal や Other のページでがんの疾患を説明する場合，公的な機関による情報である Authorized のウェブページを参照して記述される場合がある(Fig.3)．この場合も，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．以上のことから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，本研究では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ことを目的とする事とした．

4. 提案手法

4.1. 基本的なアイデア

がん情報では，３つ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．我々は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ていること．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

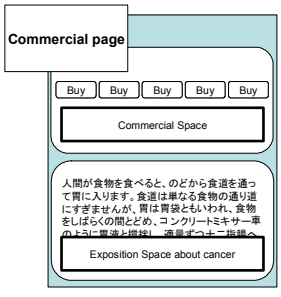

Fig.2: Other(Commercial)のウェブページ例

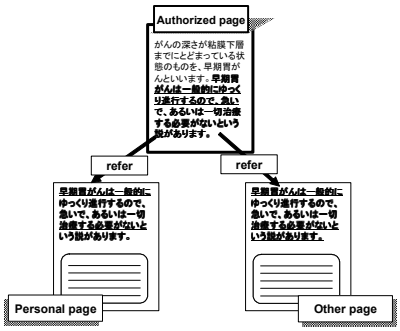

Fig.3: Authorized のページを参照する例

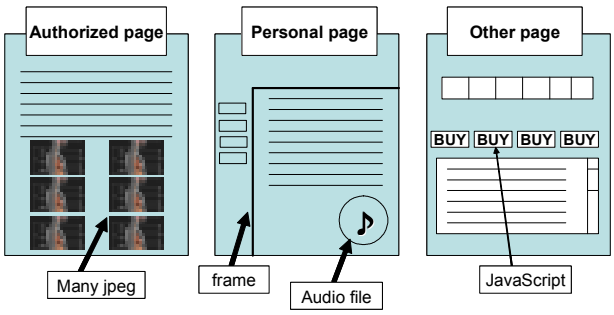

Fig.4: 基本的なアイデアの例

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．

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．そこで我々はウェブページには表面的には現れないが，ウェブページの内容を表す head 要素に着目した．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，ディスクリプションなどが記述される．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は認知されないが，キーワードやディスクリプション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．以降本研究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．

4.2. 分類に用いる素性

以上の検討から本研究では，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（約 19 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．

4.2.1. 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松本らによる ChaSen + ipadic[8]を使用した．なお，専門用語が認識できるように，ipadic には木村・中川[6]が作成したがん専門用語集 3315 語と医学専門用語約 59533 万語を追加した．専門用語比の式を(1)に示す．

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

f(Ti)は各ウェブページ i において出現するすべての専門用語の頻度である．f(Wj)は各ウェブページ i において出現する全ての名詞と専門用語の頻度である．

4.2.2. 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ウェブページの総量やイメージ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，head 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 ファイル総容量(html_size)
2. html ファイル総数(html_number)
3. jpeg 総容量(jpg_size)
4. jpeg 総数(jpg_number)
5. gif 総容量(gif_size)
6. gif 総数(gif_number)
7. png 総容量(png_size)
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表すものである[9]．head 要素には title 要素を子要素として必ず含む．その他に，文章の内容に関する 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 とは meta タグの一要素であり，ウェブページの作成者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 とは 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 要素の中にはページ作成者によって子要素を任意の数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)が使用されているか(css)
3. ページ内で flash が使用されているか(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[1]．
4. ページ内で audio ファイルが使用されているか(audio)
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置かれている index.html であれば，深さを 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’’, “ac.jp”のことである．

Table.1: Summary of the Dataset

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

5. 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性により，分類精度が変動する事が知られている．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，最悪の場合，分類精度が低下する．そこで，前項で列挙した変数のうち，CL-Score を従属変量として一般線型モデル（GLM）を適応してこれら諸値から変数選択を行い，分類精度を高めるものを検索することとした．

5.1. 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 URL(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 URL 数を Table.1 に示した．

5.2. ウェブ形態素性諸値の検討

675URL を対象として，前項で述べた，専門用語

[1] IT 用語辞典 http://e-words.jp/

Table.2: 15 Variables of URL Structural Data and “Technique Rate” at 675 Cancer URL

N of Cases	675	
	Mean	Std. Deviation
jpg_size(K Byte)	19.99	52.60
jpg_num	1.49	3.77
gif_size (K bytes)	28.74	66.01
gif_number	10.36	12.91
png_size (K Bytes)	0.44	7.15
png_number	0.11	1.19
html_size (K Bytes)	23.91	22.76
html_num	1.20	0.70
technique_rate	0.04	0.03
description_size	50.98	102.76
author_size	2.44	14.42
title_size	34.96	23.24
keywords_size	6.42	18.75
head_elements	4.46	3.19
depth	2.51	1.42

比 (techniq_rate)，ウェブ構成素性諸値 8 値（html_number, html_size, jpg_size, jpg_number, gif_size, gif_number, png_size, 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(Fig.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．

Table.3: Summary of Chi-square Test for Various four Valuables by 675 Cancer URLs

	Value	df	p-Value
javascript	11.9896	2	0.00
css	17.9828	2	0.00
flash	5.1454	4	0.27
audio	7.48503	2	0.02

Table.2 に 675URL 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

Table.4: Summary of GLM for Seeking for “Score” as Dependent Valuable by 18 Structural Valuables Selected by Stepwise (Select ; $p < 0.05$, Remove: $p > 0.1$) by 675 Cancer URLs

Model Summary

Model	R	R Square	Adj. R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1: a	0.30	0.09	0.09	0.73	0.09	65.83	1	671	2.35E-15
2: b	0.36	0.13	0.13	0.72	0.04	32.39	1	670	1.885E-08
3: c	0.40	0.16	0.15	0.71	0.02	19.79	1	669	1.012E-05
4: d	0.41	0.17	0.16	0.70	0.01	10.69	1	668	0.0011342
5: e	0.43	0.19	0.18	0.70	0.02	15.20	1	667	0.0001067
6: f	0.44	0.20	0.19	0.69	0.01	6.87	1	666	0.008982

a Predictors: (Constant), VAR00012

b Predictors: (Constant), VAR00012, VAR00013

c Predictors: (Constant), VAR00012, VAR00013, VAR00018

d Predictors: (Constant), VAR00012, VAR00013, VAR00018, VAR00014

e Predictors: (Constant), VAR00012, VAR00013, VAR00018, VAR00014, VAR00004

f Predictors: (Constant), VAR00012, VAR00013, VAR00018, VAR00014, VAR00004, VAR00019

Dependent Variable: Score

VAR00012=technic_rate, VAR00013=description_size, VAR00018=depth, VAR00014=author_size, VAR00004=jpg_size, VAR00019=javascript

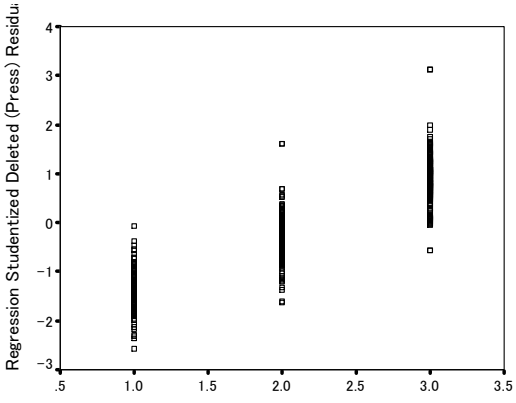

Fig. 5: Scatter plot of Predicted Value by GLM- Model “6” for Estimation for “Score”

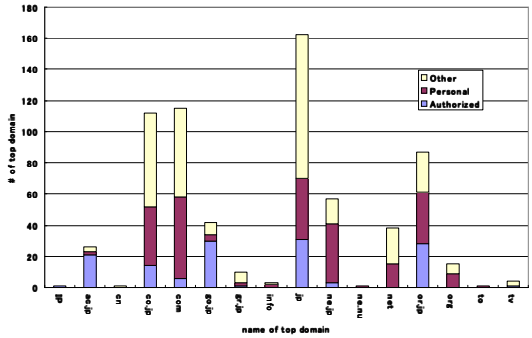

Fig.6: 各クラスにおけるそれぞれのトップドメインの出現頻度

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 Table 3 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．

Table.4 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 6 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 Fig.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（Fig.6 に示す）を分類のためのベクトル化変数として用いる事とした．

6. 評価実験

6.1. 実験に用いた 2 種類の素性のセット

本研究で提案した素性の有用性を検討するために，以下の 2 種類の素性のセットを作成し，分類器にそれぞれの素性セットを与えた結果得た，分類精度を考察する．

- feature set 1:
  GLM によって選択されたウェブ形態情報 6 個とドメイン情報（計 7 個の素性）．重回帰分析を行い，一般線形モデルから選択された計 6 個の変数(jpg_size, techniq_rate, description_size, author_size,depth, javascript)と top_domain を素性セットとしたもの．
- feature set 2:
  ウェブ形態情報(計 19 個の素性). 本研究で提案した計 19 個の素性すべてを分類器に与える．

6.2. 評価

5.1 節で説明した計 675 ページのデータセットと，前節で説明した 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[10] と C4.5[11]の二つの分類器を使用し，10 交差検定法を使用した．

まず,各素性セットの素性の数を Table.5 に示す．提案した素性はそれぞれ，7 個，19 個であり，少数の素性セットを用いて分類する．それぞれの分類器の分類結果を Table.6 に示す．この表の値はそれぞれカテゴリでの F-measure の平均値を示したものである．この結果からわかるように，SVM の場合は提案した 19 個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．

Table.5 各素性セットの素性の数

	# of attributes
feature set 1	7
feature set 2	19

Table.6: GLM によって選択された素性で分類した結果と，提案した素性凡てを用いて分類した結果

	C4.5	SVM
feature set 1	0.64	0.55
feature set 2	0.62	0.6

7. 実験結果の考察

前節の実験で GLM によって選択された 7 値の素性から構成される素性セットと，本研究で提案した 19 個の素性すべてを用いた素性セットを用いて分類実験をした結果を示した．実験結果から C4.5 の場合は，本研究で提案した 19 値の素性セットを用いて分類した結果よりも, GLM により選択された 7 値の素性セットを用いて分類したほうが良い分類精度を得た．C4.5 においては 2 種類の素性セット両方とも 6 割以上の分類精度を得たことから，分類に対するウェブ形態素性の有用性があることが示唆された．SVM においては，7 値の素性セットよりも, 19 値の素性セットを用いた分類のほうが良い精度を得た．

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．

8. まとめ

言語以外のウェブページの分類に有効な素性を発見し，その有用性を検証するために，19 種類のウェブの形態的な素性を提案し，それに対してウェブ構造素性諸値の統計的検討を行った．その結果，jpg_size, techniq_rate, description_size, author_size, depth, javascript の 6 値が分類に有効であることが示唆された．そして，この 6 値の素性にドメインの素性を加えた計 7 値の素性セットと，提案した 18 値の素性すべてにドメインの素性を加えた計 19 値の素性セットを，C4.5，SVM の分類器を用いて分類し，比較検討を行った．実験結果から，7 値の素性は SVM においては精度が 19 値の素性よりも分類精度が劣った. しかし C4.5 では文書分類においてはとても少ない 7 値の素性で約 64%の F-measure を得た．このことから，ウェブ形態情報はとても少ない素性でウェブページを 6 割程度精度で分類できることを示した．

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 SPECIAL HOME PAGE, http://www.nhk.or.jp/special/libraly/06/l0001/l0107.html

[2] 長沼，速水，“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

[3] 日本インターネット医療協議会, http://www.jima.or.jp/

[4] 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, DEWS2006 Proceedings, 1b-i10, 2006

[5] 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－, DEWS2006 Proceedings, 1b-i9, 2006

[6] 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, NLP2006 Proceedings, 2006

[7] 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, pp21-24, 2006

[8] 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

[9] W3C Technical Reports and Publications, http://www.w3.org/TR/

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

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