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－

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 Web 上のがん情報には有用なサイトが多く存在するが，専門医によって記述された文章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．今回，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思われる．本稿では闘病記や医師個人サイトを抽出する手法に関して，医師によって分類された教師データを元に Bayesian classifier を実装し，Web 上のがん情報を分類した．分類した結果，がん情報特有に現れる言語空間の調査をし，がん患者のための情報コミュニティ形成のための道具の選択と空間の定義を行った．

キーワード テキストマイニング，情報検索，テキスト分類，がん情報

1. はじめに

近年，Web 上の医療情報の増加は著しく，情報検索に困難が強いられるようになってきた．昨今 NHK[1]でも報道されたように，特にがん情報については客観的で正確な情報が少ないことから治療の選択に迷い，がん難民という言葉さえ生まれている．中川らは[2]，がん情報提供量は日米間で大きな開きがある．わが国の特異性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ころにある（Table1）．むしろ難解な医学用語で行われる専門的な説明よりも，それぞれの患者の病状に近く，連帯感を持てる患者個人の発信内容の方が医学的な正確さに優先する可能性を示唆した．特にこれら個人的な情報発信は公的な責任を前提としないため，商用目的の誘導や，いわゆる「荒らし」による標的となりやすい．せっかく検索してダウンロードできても商品販売や“荒れた”ページであっては情報取得自体の意欲が低下するという問題がある．逆に Stop Word による選択を行った場合，掲示板等にたった一つの書き込みがあっただけで除外してしまう場合や，特定語を Seek され別の手法を作られることも少なくない．いわゆるドメイン別（公的情報発信の場合は ac, go と一部の org ドメインで判別可能）分析でのフィルタによるようなものは難解な専門用語が並ぶため有効な情報を理解することが困難である．また，医師個人によるがん情報の Web ページの多くを占める闘病記や体験記に関しては，日記形式で綴られているものが多いため，がんの進行状況や，病状などが無秩序に存在するため，これを整理しなくてはならない．

Table 1: Proportion of Contents Distributors for Bile Cancer between US and Japan at Top 100 Hits.

Contents Distributor	US	Japan
Hospitals and Universities	4	16
Organized Institutions	50	27
Patients and Families	0	2
Individuals (Including M.D.)	6	17
Cancer Information Distributor	12	4
Medical Portal site	4	6
Publisher	19	17
Others	1	5

これら分別が困難な個人的情報発信を行っている URL をできるだけ効率的に患者に提供し，同病・同病期の患者コミュニティを形成しやすくするのが今回の研究の目的である．最近 P2P ネットワークで試行されている RSS[3]や奥村らの Burst 検出[4]等の関連研究もあるが，今回我々はこれら一般概念の分析ではなくがん情報の特殊性に特化しモデル化することと，教師データに対する正解率を目的として,「がん分野の個人による情報発信」のカテゴリおよび他のカテゴリを選択するための言語空間の推定を行う．

以下，本稿では，現状の Web 上のがん情報の言語空間の推定を調査に基づいて行う．次にコンテンツ空間の分類をする．ここで Web 上のがん情報を分類する手法を示す．3 節では 2 節で実装した Naïve Bayesian classifier の実験結果を示す．

Table.2 CII(cancer Information Index)
Definition of Category of Cancer Information

1：Authorized Information 学会、学術研究機関からで Peer reviewを行っていると思われる情報
2：Unauthorized Infor Peer Review なしの情報 闘病記、医師個人、患者コミュニティ などの情報も含める
3：Media Information Portal、書籍など
4：Other Information 商用の宣伝など
5：Noise 検索目的に合わないもの 検索語を含まないもの

4 節では実験で得られたデータから Web 上のがん情報の考察を述べる.

2. コンテンツ空間の分類

2.1. 分類

Web ページの分類は，中川ら[2]の研究によって定義された CII (Cancer Information Index)を適用する．これはがん情報を Web ページの内容に基づいて 5 つのカテゴリに分類する定義である．CII を Table.2 に示す.

2.2. 全体の流れ

Web ページを 5 つのカテゴリに分類するために，我々はベイズの定理に基づいた Naïve Bayesian classifier[5]を実装した．近年，文章の分類に関しては SVM[6]といった手法のほうが多く用いられているが，Naïve Bayesian classifier を選んだ理由は，本稿では Web 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，シンプルでわかりやすい上に分類の正解率も高い Naïve Bayesian classifier を分類器として選択した．また，単語の頻度情報が得られるために今後の研究に役立てることも念頭において実装することにした.

実装する Naïve Bayesian classifier の全体の処理の流れと実験の流れを Fig.1 に示す．Naïve Bayesian classifier の処理を 2Step に分けて説明する. 2.4 節で詳しく述べるが，まず，Step1 であらかじめ人間によって分類された Web 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，Step2 に処理が移り，Step1 で学習したトレーニングデータを用いて分類器の正解率を測る．テストデータは検索エンジン Yahoo! JAPAN でそれぞれ胃がん，大腸がん，子宮がん，乳がん，白血病(以降省略してそれぞれを，sc,colon,uc,lc,lekumeia と呼ぶことがある)を検索語として検索した結果得られた上位 30 件を

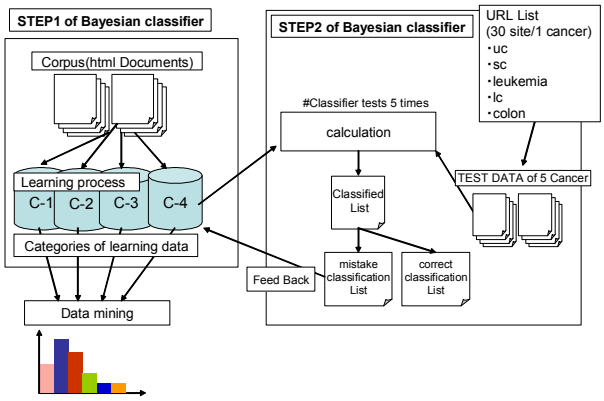

Fig.1 Overview of Naïve Bayesian classifier

医師によって分類された結果をテストデータとした．テストデータを Naïve Bayesian classifier にかけた結果，判断を誤ったものに関しては Step1 で作成したトレーニングデータにその情報をフィードバックし，トレーニングデータを訓練しなおす．このように同じテストデータを用いて，同じ処理を 5 回繰り返し，トレーニングデータの充実と言語情報の獲得を行う．そして，4 節でトレーニングデータから得られるがん情報の言語情報を分析した結果を示す.

2.3. 初期教師データ

Step1 で使用する初期の教師データは医師の監査の下で Yahoo! JAPAN の癌カテゴリから計 31 サイトを選出し，Table.2(CII)の定義に従いカテゴリに分類した．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 Web ページをダウンロードした. Category1 の教師データとなるサイトは本稿の実験では国立がんセンター[7]の Web ページをすべてダウンロードし，それのみを教師データとした．Category1 の教師データに国立がんセンターのみの Web ページを使用した根拠は，木村ら[8]が示したように，国立がんセンターのがんの解説ページは，がんに関する文章で標準的に使用する単語を多く含むため妥当であると考えたからである．Category2 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 Web ページが主な内容である. Category3 はがん情報の書籍情報や，がん情報のポータルのサイトを選出した．Category4 は，がんの漢方薬販売の Web ページを主に選出した．Web 上に存在するがんに関する販売目的の Web ページの多くは漢方に関するものであるため，初期教師データは漢方販売のページのみに絞った．そのほかの Category4 に分類されるべき Web ページは，Step2 のテストをするときに分類を誤ったものを訓練しなおすようにした. Category5 は Web ページの header

や footer に検索語を含むものであり，その Web ページから文章を抽出すると，”がん”という単語が出現しないものである．よって計算コストを軽減させるために Naïve Bayesian classifier で分類せずに，Web ページの本文中に”がん”という単語が出現しない場合は Categoy5 に分類するフィルタを作成した．最終的に得られた個々の Web ページを Naïve Bayesian classifier で処理するために，html データから html タグを外し文章のみを抽出した．

2.4. Naïve Bayesian classifier の実装

2.4.1. Step1

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 Bayesian classifier に学習させる Step である．本稿で作成するトレーニングデータは，Web ページから抽出された文章を教師データとし，それに対して chasen[9]を用いて形態素解析した結果得られた単語の頻度をカウントする．そのデータをトレーニングデータとして各々のカテゴリに作成する．ただし，カウントする単語は名詞に限定し，かつすべての Web ページは，がんに関するものなので単独で出現する”がん”, ”ガン”,”癌”という単語はすべての Web ページに含まれるため分類の判断に影響を及ぼさないであろうと考え，カウントから除外した．ただし，”胃がん”,”抗がん剤”など，”がん”という単語が複合語で現れる単語に関してはカウントするようにした．このようにそれぞれのカテゴリにおいて，教師データの中に出現した単語をカウントし，トレーニングデータを作成する．分類の判断は文章の文脈や単語の出現箇所を考慮せずに単語の出現数のみを考慮した単純なモデルである．

2.4.2. Step2

Step1 でトレーニングデータを作成したら，Step2 の処理に移る．前述したように，本稿で実装した分類器はベイズの定理に従った Naïve Bayesian classifier である．ここの処理は[10][11]の実装を参照して作成した．Step2 では，どのカテゴリに属するのかを判断させるために新しく得られた各々の Web ページを Step1 と同じように文章を抽出し，そのデータに対して chasen を用いて形態素解析をして形態素に分割する．そしてその Web ページの出現単語数をカウントする．

各カテゴリを$\{c_1, c_2, \cdots, c_5\}$とおき，各文章を$\{d_1, d_2, \cdots, d_j\}$とおく．そして，$d_i$に出現する単語を$\{w_1, w_2, \cdots, w_k\}$とおき，読み込まれた Web ページ d_j は事後確率$P(c_i|d_j)$ を最大化するような$\hat{c}$ が求められる．$\hat{c}$ は次式で求めることができる．

$$\hat{c} = argmax_{ci} P(c_i|d_j) \quad \cdots (1)$$
$$= argmax_{ci} P(c_i|w_1, \cdots, w_n) \quad \cdots (2)$$
$$= argmax_{ci} P(w_1, \cdots, w_n|c_i) P(c_i) \quad \cdots (3)$$

そして，Naïve Bayesian classifier の定義に従って各カテゴリにおいて単語は独立に生起すると仮定し,文章の分類は次式で求められる．

$$\hat{c} = argmax_{ci} P(c_i) \prod_{k=1}^{n} P(w_k|c_i) \quad \cdots (4)$$

(4)式で，$P(c_i)$ は(c_i に含まれる Web ページ数 / すべての Web ページ数)で求める．また，c_i に出現する総単語数をN_i,c_i において w_k が出現する回数を F_{ik} とおき，

$$P(w_k|c_i) = F_{ik} / N_i \quad \cdots (5)$$

と計算する．

以上の計算がオリジナルの Naïve Bayesian classifier の主な計算方式なのだが，本稿での分類の対象ドメインはがん情報であるために，新しい Web 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に，もし一単語でも F_{ik} が 0 になった時は確率が 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 0 頻度の問題を解消するために，出現するすべての単語の頻度に 0.5 をあらかじめ足し，すべての単語の異なり数を V とおき，(6)のように定義する．

$$P(w_k|c_i) = (F_{ik}+0.5) / (N_i+0.5\,V) \quad \cdots (6)$$

読み込まれた Web ページに出現する単語が教師データの中に存在しなかったときは，

$$P(w_k|c_i) = 0.5 / (N_i+0.5\,V) \quad \cdots (7)$$

となる．

2.4.3. 計算式の修正

Naïve Bayesian classifier は基本的には初期の教師データを学習しただけでは分類の正解率があまり高くなく，適度な学習を繰り返すことで正解率が上がる．学習を繰り返すことで，トレーニングデータが増大し計算的なコストは高まるが，分類精度の向上が望める．後に詳しく説明するが，本稿ではまず初期教師データのみで Naïve Bayesian classifier の正解率を評価し，

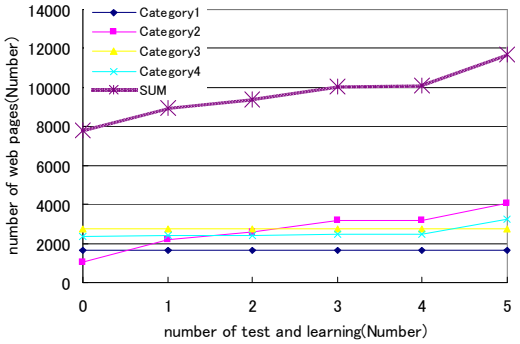

Fig.2 number of training times and number of selected Web pages by Naïve Bayesian classifier

誤って分類したものは正解のカテゴリの教師データに追加し，再度評価するといった処理を 5 回繰り返す．これを繰り返すことによって教師データとなる Web ページは線形に増加する様子を Fig.2 に示す．横軸にはテストの回数(テスト + 学習)を，縦軸にはトレーニングデータの中に含まれる Web ページ数をプロットした．5 回の学習を終えたときでは初期の教師データの約 1.5 倍となっている．正解率の向上を望み学習を繰り返すごとに計算コストは増加していく．特に本稿が取り扱うような対象データが Web コンテンツといったドメインの研究では文章数が莫大であるために Step2 の(6)(7)で定義した式では分母が過大化する上に，積を取るために多くは確率が 0 になってしまう．

そこで莫大な量の文章を処理する時でも計算が可能となるように(4)を修正した．Step2 では計算で積を取っているが，対数を計算し，それを最大にするような $\hat{c}$ を選択するようにした式を(8)(9)に示す．

$$\hat{c} = argmax_{c_i} \log\left(P(c_i)\prod_{k=1}^{n} P(w_k|c_i)\right) \quad \cdots (8)$$

上の式を展開すると，

$$\hat{c} = argmax_{c_i}\left(\log\left(P(c_i)\right)+\sum \log\left(P(w_k|c_i)\right)\right) \quad \cdots (9)$$

となる．本稿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，(8)(9)を適用した．

3. 実験と結果

3.1. 解析不可能な Web ページ

これまでに説明してきたように Naïve Bayesian classifier は Web ページの言語情報に依存して判断する．本研究で分類しようとしている文章は Web ページであるために，言語情報がごくわずかで，ページ上の多くが画像データの場合がある．特に Web サイトの Top ページの場合は文章ではなく，そのサイトに存在するコンテンツ名のリストのみが記述されている場合がある．また，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，言語情報量が 150byte に満たないページに関しては本稿の分類器では判断に十分な情報量ではないとみなしテストデータから対象外とした．Category5 に関しては，「がん」という言葉が存在しないページを分類される．よって，本稿で実装した Naïve Bayesian classifier は上記の条件を満たした Web ページを Category1 から Category4 のいずれかに分類する．

3.2. 評価

ここでは，”子宮がん”,”胃がん”,”白血病”,”乳がん”,”大腸がん”をそれぞれ検索語として Yahoo! JAPAN で検索した結果得られた上位 30 サイト(計 150 サイト)を医師によって分類して，それらの Web ページをテストデータとして実験した．最初に初期の教師データのみで学習したトレーニングデータを使用した Naïve Bayesian classifier の判断結果を得た．そして誤って判断した Web ページの単語の頻度情報を正しいカテゴリのトレーニングデータに追加して再度同じテストデータを用いてその処理を 5 回繰り返した．

実験結果を Fig.3,Table.3 に示す．Fig.3 の横軸はテストの回数を示したものであり，縦軸は各テストでの

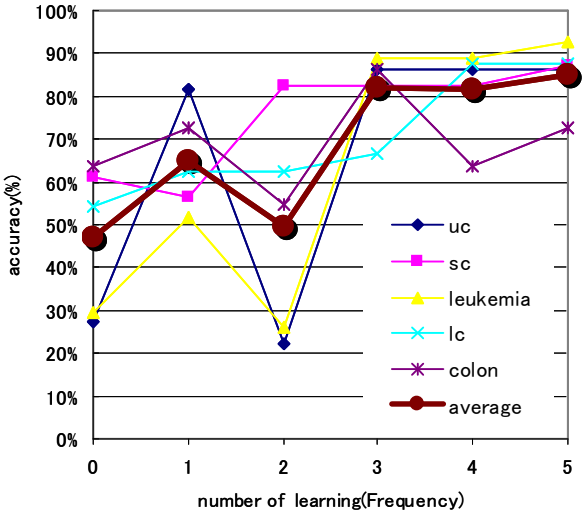

Fig.3 Result of the Naive Bayesian classifier accuracy

Table.3 Result of the Naïve Bayesian classifier accuracy. The number means learned frequencies

	0	1	2	3	4	5
uc	27.3%	81.8%	22.3%	86.4%	86.4%	86.4%
sc	60.9%	56.5%	82.6%	82.6%	82.6%	87.0%
leukemia	29.6%	51.9%	25.9%	88.9%	88.9%	92.6%
lc	54.2%	62.5%	62.5%	66.7%	87.5%	87.5%
colon	63.6%	72.7%	54.5%	86.4%	63.6%	72.7%
average	47.1%	65.1%	49.6%	82.2%	81.8%	85.2%

正解率を示したものである. Fig.3 が示すように教師データを増加したからといって必ず分類の正解率が向上するわけではない．Table.3 の行はテストの回数を，列は各がんの疾患名を示し Fig.3 の正解率の数値を示したものである．この結果が示すように，学習を繰り返し行うことによって高い正解率を得られるようになる．最後に，学習には使用しなかった「卵巣がん」のテストデータを使用して, 同じように 30 サイトで分類の実験をしたところ，83.3%の正解率を得ることができた．

4. Web におけるがん情報の言語情報の考察

本稿の研究で Naïve Bayesian classifier を実装することで，Web 上のがん情報の各カテゴリにおいて出現する単語の頻度情報を得た．この情報を分析した結果を論ずる．

4.1. 言語空間の考察

本稿で実装した分類器は 3 節でも説明したように単語の頻度情報によって分類を判断する．そこで，Category1,Category2,Category3,Category4(以下それぞれ C-1,C-2,C-3,C-4 と呼ぶ)のトレーニングデータの単語頻度情報を比較した結果を Fig.4 に示す．横軸にはそのカテゴリに出現した単語を頻度の小さい順に並べ，縦軸には単語の distance をプロットした．例えば 1and234 であったら，まず C-2,C-3,C-4 のトレーニングデータを元に，出現単語頻度をそれぞれの単語に対して足していき和集合を作成し，新たに C-2,C-3,C-4 を一つの集合としたカテゴリ(C-234 と呼ぶ)を作成する．なお，最終的に作成した和集合のそれぞれの単語の頻度は 3 で割わり，平均を取ったものである．

そして，C-234 の単語頻度情報を得たら，C-1 のそれぞれの単語頻度から C-234 での単語頻度の差を取ったものが 1and234 である．同じように C-2 と C-134，C-3 と C-124 そして C-4 と C-123 の頻度情報の差を取った．その差を縦軸の distance とした．

具体的には，C-1 には「研究」という単語が 9977 回出現する. それに対して, C-234 では 1506 回出現する．この差を取ると 8471 回となり，C-234 に対して C-1 では「研究」という単語が 8471 回多く出現しており，これは C-1 に特徴的に現れる単語だとわかる．

逆に，C-1 で「漢方」という単語は 4 回出現しているのに対して，C-234 では 8794.5 回出現している．差を取ると-8790.5 回となる．つまり，「漢方」という単語は C-1 にはほとんど出現しない単語であり，「漢方」が出現したら C-1 の Web ページではない可能性が高いことを示唆している．

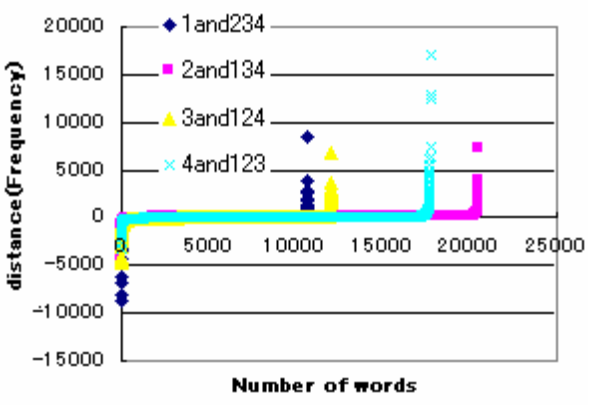

Fig.4 word specificity

ここで注目するべき点は，distance が 0 の単語が多く存在していることである. distance が 0 ということは，つまりその単語は Web ページを分類する際に影響していないことを意味する．よって，分類をする際に手がかりとしようとしている出現単語の頻度情報の多くは distance が小さく似通っていることがわかる．そして，あるポイントを境に急激に distance の差が開いていることがわかる．この考察はがん情報を分類するのに大いに役立つのではないかと考えている. なぜなら，本稿で実装した Naïve Bayesian classifier は計算コストが高いのだが，上で述べたように急激に差が開くポイントのみを分類の情報として使用すれば，計算コストを大きく軽減できるのではないかという仮説を立てており，今後の研究に役立てたいと考えているからである．

4.2. 各カテゴリの言語的特長

ここでは先ほど示した Fig.4 の言語情報の特徴を詳しく考察して論ずる．Table.4 は Fig.4 で用いた元データの distance の上位 10 単語と下位 10 単語を示したものである．1and234 の表の考察を述べる．C-1 は 2 節で述べたように Peer Review された Web ページである．よって，特徴的に現れる単語が「研究」であったと推測される．その他に目立つ点は「一覧，目次，内容，更新」といった，Web サイトの構造及び内容を解説する単語が多く存在していることがわかる．おそらくこれは，国立がんセンターのような研究機関はがん情報を頻繁に更新しており，かつ，整理してそれを表示しようとする試みから，他のカテゴリの文章よりも特徴的に現れたのではないかと考えている．

逆にマイナスの distance を考察してみると，「漢方」という単語は C-1 ではほとんど使用していない．個人や，企業が発信する情報と学術研究機関が発信する情報の大きな違いはここにある．そして，この情報の違いのギャップこそが，がん情報検索者を困惑させる大

Table.4 characteristic words of each category

1and234

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

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

3and124

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

漢方	17095	研究	−3304
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

きな問題の一つであると我々は考えている．例えば，がんの治療に関して検索をした時に C-1 では，専門的な最新医療に関しての情報を発信しているのに対して，C-234 では体験談による医療の解説もあれば，漢方で治ると述べているページ(特に C-4 で)もある．こういった現状ではがん情報の知識が少ない検索者であれば何が正しいのか判断できず，困惑してしまう可能性が高いことが推測される．この問題は命に関わることであり深刻である．そして，その他のマイナスの distance がある単語の特徴としては，「相談」，「私」といった単語が特徴的である．C-1 は研究機関の Web サイトであるために，Web 上から「相談」することはめったにない．「相談」という言葉は経験的にがんの医薬品会社が客に対して相談を促す場合に使う時と，がん患者による闘病記で患者が医師に相談した時に使用した記録が多い．こういった情報こそがん患者にとって必要な情報なのではないかと我々は考えている．C-1 では，各がんの概要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章は難解である．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様々な違いがある．そういった「相談」したいような内容は C-1 には少ないことを示唆している．C-1 の情報だけでは足りないような付加的な情報を C-2 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得ることによって検索者は求めていた答えに近づけるのではないかと考えている．次に，「私」という単語は一人称で用いる単語であり，C-1 には現れることは少ない．これは闘病記や体験記に特徴的に使われる単語である．

その他のカテゴリの特徴的な言語的情報を述べると，2and134 では「先生」という単語が特徴的に現れている．医師によって記述された Web ページでは「先生」という単語よりも，「医師」という単語が一般的に使用される．実際に C-1 のトレーニングデータを考察すると「先生」という単語は 168 回使用されているのに対し，「医師」という言葉は 1469 回使用されている．これは一例過ぎないのだが，医師が記述するがんの Web ページと医師以外が記述する Web ページ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．特に専門的な用語に関しては多く違いが見られる．

5. おわりに

今回，Web 上のがん情報を整理するために Naïve Bayesian classifier を実装した結果，以下のことが分かった．

- Web 上のがん情報は言語情報に特徴を持ち，Naïve Bayesian classifier のように言語情報を元に自動的に分類するような手法でも分類が可能である．
- Fig.4 で示したように，CII の各カテゴリに属する Web ページに出現する言語空間は多くが似通っているが，各カテゴリを認識するための特徴を現す言語が存在する．
- カテゴリの特徴を現す単語を特定できたことから，Web ページの分類に使用する素性を減少できる可能性がある．

分類に用いたトレーニングデータの言語空間を単語単位で分析した結果，以下のことが分かった．

- 情報の発信源によって使用される単語に相違が見られ，専門医，がん体験者，販売業者間では使用される言語空間に違いがある．
- 各カテゴリでは特有に使われる固有名詞が存在し（例えば C-4 であったら漢方などの固有名詞が多く使われるように），カテゴリを識別するのに固有名詞の情報も活用できる．
- C-2 で「私」という単語が多く使われていたように，カテゴリによって人称の使い方に違いがある．

これらの得られた結論は，Web コミュニティ形成の研究を進めるのに有用なデータとなる．今後は得られたがん情報の特徴を用いて，具体的な Web コミュニティ形成のための検討をしていきたい．

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，北陸先端科学技術大学院大学白井清昭助教授，情報通信研究機構竹内友木子氏，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 17 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] 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

[3] Resource Description Framework (RDF)/ W3C Semantic Web Activity http://www.w3.org/RDF/

[4] 藤木稔明,南野朋之,鈴木泰裕,奥村 学,"document stream における burst の発見",情報処理学会研究報告, 2004-NL-160, pp.85-92.

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

[6] Susan Dumais，Hao Chen, ”Hierarchical classification of Web content”, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR2000),pp.256-263,Athens,Greece,July2000.

[7] 国立がんセンター http://www.ncc.go.jp/jp/

[8] 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”,言語処理学会全国大会 2006(投稿中)

[9] 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

[10] 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.

[11] 阿部倫子，田中久美子，中川裕志,"コメントを用いた映画の分類" 情報処理学会 NL 研究会 NL-150, pp.105-110, 2002 年 7 月