Web がん情報評価のための単語集合の作成と検討

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

†情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

‡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 1-2-3

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

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

あらまし Web 上で提供されているがん関連情報適正化を目的として、医師による評価値を教師データとした外的基準（CII）をもとに専門用語出現頻度、数種の Web マイニングデータによる解析、あわせてベイズ法による分類アルゴリズムの作成を行った。その結果、ベイズ法では accuracy が 80%であり本分類が良く適応することが示された。分類後の各グループにおける頻出語彙は、医学的な用語ではなく、「先生」「私」等、発言者の語調に起因すると思われるものが多く問題であることが示された．今回、解析に用いる用語集合を定義しなおし、妥当な外的基準を与えることを目的として専門用語辞書作成方法について検討した．

キーワード ウェブマイニング，情報検索，文書分類

1. はじめに

1.1 ウェブがん情報の質に関する検討

治療法の確立されていない疾患であるがん[1]を発病したがん患者にとって，日々刻々更新される最新のがん情報を的確に得ることは延命や治癒のために，手術，内服薬に匹敵する第三の薬である．医療現場で医師から与えられる情報の正確さと根拠について，より的確な情報を早期に得，より的確な医療を受けることが予後（生存期間）の延長や余命の中での生活の質向上に直結するからである[1]． インターネットによるがんに関する情報発信は活発化し情報の量が増加してきているが，中川・木村は，数種のがんについて，わが国で発信されているこの分野のコンテンツを分類し，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を報告した[2]． 専門性の高い研究指向のコンテンツ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る．むしろ個人を対象とした個人（情報ボランティア）による情報発信からの情報がより患者のニーズに近い情報を与える可能性があることを示唆した[3]．これらから、Yahoo!や Google 等の汎用される検索エンジンにより提示される順番は、Fig.1, Fig.2 に示すように、図中の A-type のように上位ほどノイズが少ないのが理想であり、例えば Fig.2 の白血病などではこのパターンを示すが、逆に CC（大腸がん）のように上位ほどノイズの多い傾向を示す場合など一定しなかった

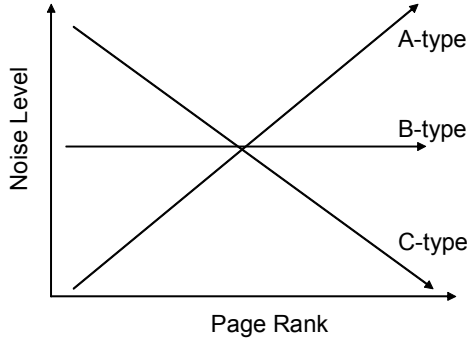

Fig.1: Schema for Relations between Listed Page Rank and Noises at Usual Search Engines

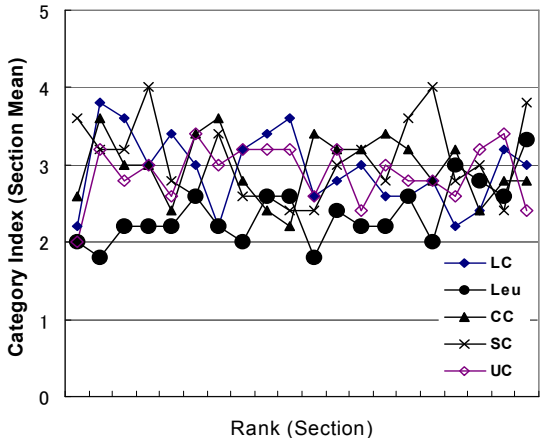

Fig.2: Section Mean (/5URLs) of Category Index by Listed Ranks of URLs

注：疾患名で「がん」を「癌」あるいは「ガン」と表記する場合もあるが、本研究では国立がんセンターの慣例に習い、「がん」で統一する。尚検索エンジン等で URL を取得する場合には「がん、癌、ガン」の 3 通りで検索し OR を取った．

Table 1: Example of Characteristic words CII=2 (Non-authorized Contents) and CII={1,3,4}(other)

Word	Difference	Word	Difference
私	7216	研究	−4987
入院	3917	相談	−4558
病院	3905	漢方	−3888
検査	3240	シート	−3066
自分	3214	情報	−2086
先生	2336	一覧	−2069
海外	1875	抗がん剤	−2062
手術	1871	内容	−2034
これ	1816	必須	−1739
人	1805	薬局	−1599

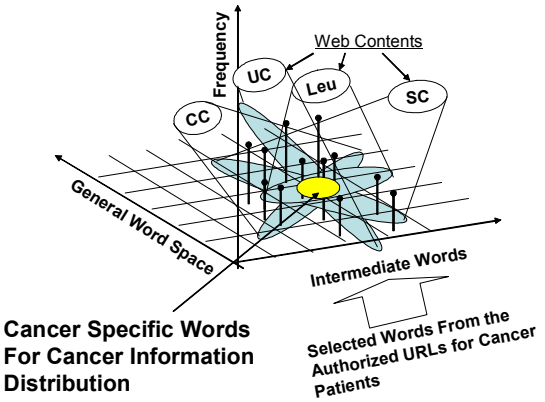

Fig. 3: Assumed model of relations among various word spaces appearing in Cancer Webs

（C-type に相当）[4]. 予備実験として一般用語辞書(i-dic)を用い URL 毎の単語出現頻度を元に数値化、機械分類を行った[5]. Naïve Bayesian classifier での的中度は 80%以上であり、教師データに対する分類精度は高かった.

1.2 がん用語集合の必要性と研究の目的

分類の質を検討するために、分別されたグループ毎に頻出する単語を Table 1 に示した.（頻出した語が本分類において寄与が高かったと考えられる）が医学用語ではなく、「先生」「私」等の名詞であり、分別されたグループが必ずしもそれぞれに特徴的な医学的用語の軸によって分類されたものではないことがわかる. これは、Fig.3 に示すように、分類において特徴量として分類器が採用した言葉集合と、医学的用語空間に違いがあり実際に分類が行なえていたとしても医学的整合性や妥当性を担保するものではない. ウェブデータにおける一般用語の生起頻度と医学用語集合の生起頻度の相関関係、共起頻度等、基礎的データを検索したが、関連研究はなかった. 特に「がん」は医学的知識の中でも、疾患概念、診断・治療ならびに予防方法がほぼ確立している高血圧、糖尿病、脳卒中などの疾患群に比べて専門性が高く、新薬や新規検査手法などの更新頻度も高いため一般の医学用語集合とも異なる. しかし、「がん」の疾患概念、診断治療方法など多岐にわたる標準語集合は未だ定義されていない.

さらに、臨床の場では、いわゆる「がん」に関する用語は大きく分けて三種類存在する. 例えば、一般には「肺がん」と言われているがんでも、医師の診断過程では「孤立性右肺陰影」や「右肺小細胞がん」あるいは「脳転移を伴う右肺扁平上皮がん」という言い方で、その患者の疾患の状態は区別され、重症度に関する医療関係者の認識も異なる. しかし、この言い方を例えばインフォームドコンセント（説明と同意）を取得する時に患者や家族に対して伝えても専門用語を並べるのみで概念が伝わったとはいえない. 高血圧や糖尿病はがんに比べて経時的変化により重症度を増すが、がんは診断時の進行度や重症度を表す語ががん専門家（医師､看護士)､医療従事者､患者それぞれに異なる. 例えば、専門家の間で言う「右肺小細胞がん」は、患者に説明するときには「右肺小細胞がん（小細胞がんという種類の右肺にできた肺がん)」となる. ウェブで行なわれている情報提供においても発信者の背景と想定される対象者によって同じ用語であるとは限らないというのが、がん情報の特殊性である.

以上のことから本研究では、「がん」に特異的な用語集合を採取作成し、がん情報コンテンツとして提供されているウェブデータを対象として、一般用語辞書、一般医学用語辞書との出現頻度ならびに諸値を比較することによって、「がん用語空間」の特性を明らかにする事とした.

2. がん用語空間のモデル化

本研究で想定するがん用語集合のモデルを Fig. 4 に示す. がん情報に関する Web コンテンツを提供する元と提供する先である Doctor (Medicine), がんに関する情報 Cancer および一般概念 General (Patient),の 3 つの連関を想定し、それぞれの言葉空間を (Wm; Words-Medicine, Wc; Words-cancer,Wg; Words-General), を想定した. 専門性の高い情報は医学専門用語によって交換されており Web で提供されていても患者にとって理解できない場合も考えられる. 上に述べたように、これら言葉群(Wm, Wc, Wg)は全く異なるのではなく、同じ事象に関する記述を学術的背景や状況によって行っているため何らかの連関性を持つ共通概念を形成している可能性もある. これら概念の抽出のために、それぞれの単語集合それぞれを実際に作成し、実際に提供されている Web コンテンツにおける生起頻度を検討することとした.

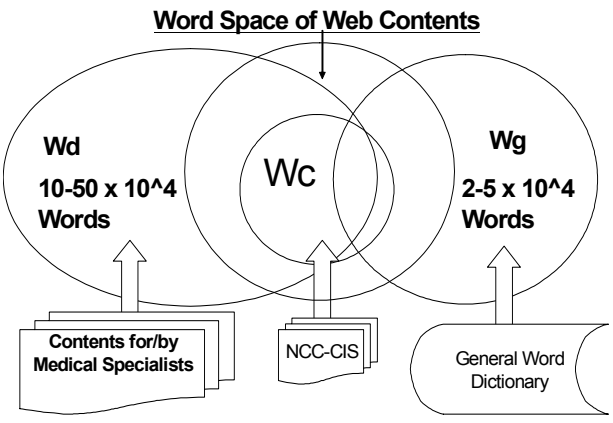

Fig.4 Assumed Relations and Scales of Each Word Set for Cancer Information Providing

2.1 想定する単語集合

本研究の対象とする単語集合は、最低単位を医学的に用いられる文節からなるものとする必要がある。そのため、例えば「転移性肺がん」という文節を従来の一般用語辞書で次のように切り出しては意味がない。

両側性肺門部リンパ節腫脹
- 両側
- 性
- 肺
- 門
- 部
- リンパ節
- 腫脹

「両側性肺門部リンパ節腫脹」それぞれ、一文節一語として頻度を算出する必要がある．そのため、未知語を採集する URL が決まっても形態素解析を行なう場合、通常の一般用語辞書を用いて既知語の「転移」「性」「肺」「がん」に分類されると結果を誤る．Wg は idic（約 7 万語）をそのまま用いることとするが、専門用語の抽出に必要なさらに長い文節からなる専門用語辞書を必要とする．専門用語を抽出するアルゴリズムも提案されており[8]、実装し抽出を試み、約 3 万語を得たが、誤抽出も相当数（約 2 割程度）あり、専門的知識を有する有資格者（医師）が直接切り出すほうが効率的であると判断した．長文節の単語からなる辞書を Fig.5 に示すように直接作成する事とした．

2.2 Wcc(がん専門用語辞書)の作成経過

国立がんセンターの Web ページ[9]の文章を元に用語辞書を作成した．がん専門用語は，国立がんセンター(NCC-CIS)で提供されている疾患別解説ページにそれぞれ出現する単語を切り出した．作成時、がんを解説している疾患数は計 54 種類あり,それぞれ手作業で専門用語（W_{c1}-W_{c54}）を切りだした．それぞれ和をとり異なり語集合を求めワードセット C1-Dic とした。

3.3 C1-Dic 作成経過

疾患別に用語集合 W_{ci} を加えたときの C1-Dic 内に存在する用語数の変化を Fig.6 に示した．縦軸は C1-Dic の語数である．がんの数を増加させてゆくとともに辞

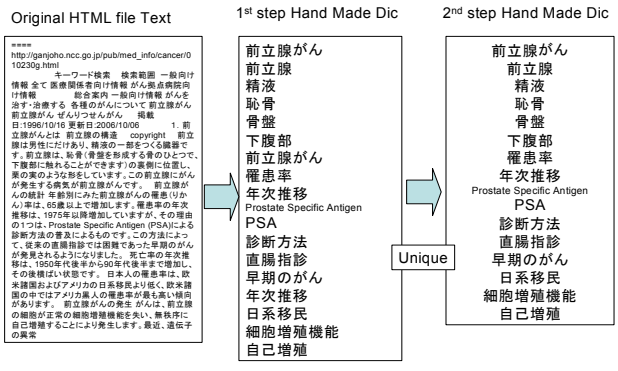

Fig. 5: Process of Making the Hand Made Dictionary from URL Contents.

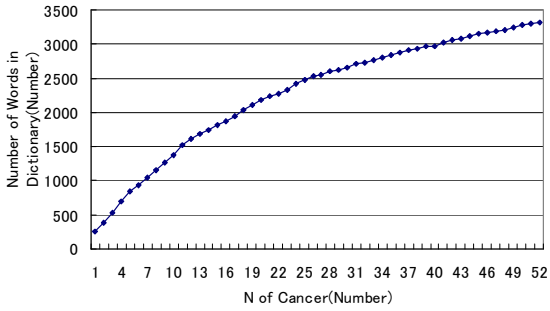

Fig.6: Number of words in C1 Dictionary

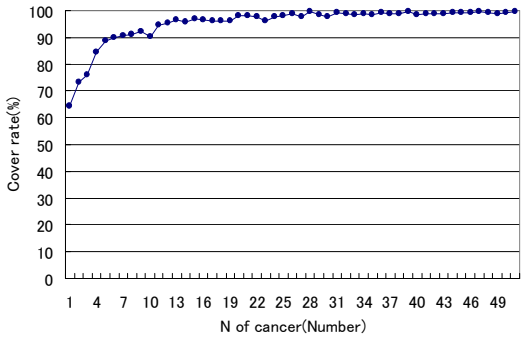

Fig.7: Cover rate of cancer dictionary by Number of Cancers

書の単語数も単調に増加するが，1 つのがんあたりの増分が減少する．計 54 種類を合わせた結果，辞書に取り入れる用語は合計 3313 語となった．Fig.9 に疾患数と Cover rate の推移を示した．これらから各疾患を解説するのに用いられる用語は多くが重複していることが示唆された．以上がん用語辞書”C1-Dic”（3313 語）を固定した．

さらに 2006 年 10 月に国立がんセンターホームページが大幅に改訂された．ダウンロードしたデータ量は合計約 250MB であった．この中から、HTML ファイルのうち、テキストデータのみを抽出し、合計 15M バイトを得た．ここから疾患数も追加され、53 から 59 となったため、本研究で検討するがんの種類も Table 2 に示す 59 種類とした.再度同様の手法で”C2-Dic”(9451 語)を作成した．

3.4 ワードセットの設定と構成の概要

C1, C2 に含まれる語彙の特性を検討するため、一般用語辞書(i-dic: 75498 語)、インタネット上で収集した複数の医学用語辞書(M-Dic: 59533 語)を加算し比較用の語集合とした．Table 3 および Fig.8 に語集合 G,M,C1,C2 におけるそれぞれの単語長の分布を示す．最長単語長は G: 13, M: 79, C1: 35, C2: 89、平均単語長 G: 2.9, M: 10.4, C1: 4.9, C2 5.9 であり、医学系 3 種類の M,C1,C2 が一般用語集合 G に比べて長い．

Fig. 8 より、G に比べ、M が 10 文字以上の単語が多い事、C1 より C2 が長い語を含んでいる事がわかる．M は複数文節からなる英単語を含んでいる事、C2 は C1 に比べ、より精密ながんに関する情報を提供する目的のために改変されたコンテンツからなっていた事か

Table 2.: Valid Cancers in this study.

胃がん	菌状息肉症	中皮腫
肺がん	形質細胞性腫瘍	聴神経鞘腫
大腸がん	原発不明がん	軟部肉腫
肝臓がん	喉頭がん	尿管がん
白血病	骨髄異形成症候群	脳腫瘍
乳がん	子宮体部がん	皮膚がん
子宮がん	子宮肉腫	非ホジキンリンパ腫
ぶどう膜悪性黒色腫	子宮頸部がん	慢性リンパ性白血病
ホジキン病	上咽頭がん	慢性骨髄性白血病
悪性リンパ腫	食道がん	慢性骨髄増殖性疾患
悪性黒色腫	神経膠腫	網膜芽細胞腫
咽頭がん	腎細胞がん	卵巣がん
陰茎がん	腎盂がん	卵巣胚細胞腫瘍
下咽頭がん	成人T細胞白血病リンパ腫	睾丸腫瘍
下垂体腺腫	精巣腫瘍	絨毛性疾患
外陰がん	前立腺がん	膀胱がん
肝細胞がん	多発性骨髄腫	膣がん
急性リンパ性白血病	胆管がん	膵がん
急性骨髄性白血病	胆嚢がん	膵内分泌腫瘍
胸腺腫	中咽頭がん	合計 44,910URLs

Table 3. Basic Statistics of G, M, C1 and C2 Word Set.

	Category of Dictionary			
	G	M	C1	C2
N of words	75498	59533	3315	9451
Averaged Length of Word	2.90	10.14	4.90	5.86
S.D. of Length of Words	1.26	7.15	2.17	4.43
Min Length	1	1	2	1
Max. Length	13	79	35	89

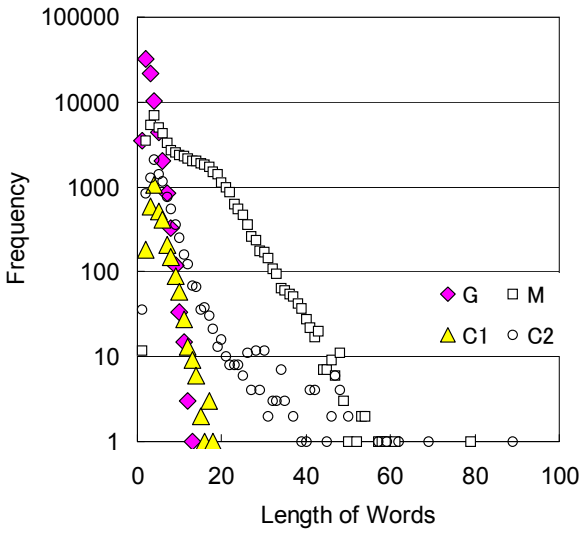

Fig.8: Comparison of G, M, C1 and C2 Word sets (Word Length – Frequencies).

ら、C2 の語数が多くより専門性の高い M に近い語彙

Table 4: Result of Survey of 59 Cancer-44477 URLs for Each Word Sets (G, M, C1, C2 and Various Duplicate Data Sets)

Category of Sets	N of Words	N of Appearance	Appearance Rate
C1	3315	2470	0.745
C2	9451	5826	0.616
M	59533	40381	0.678
G	75498	43942	0.582
C1∩C2	2494	1948	0.781
C1∩M	801	762	0.951
C1∩G	103	101	0.981
C2∩M	1974	1870	0.947
C2∩G	1046	1021	0.976
M∩G	2827	2588	0.915
C1∩C2∩M	686	657	0.958
C1∩C2∩G	66	64	0.970
C1∩M∩G	69	69	1.000
C2∩M∩G	629	626	0.995
C1∩C2∩M∩G	48	48	1.000

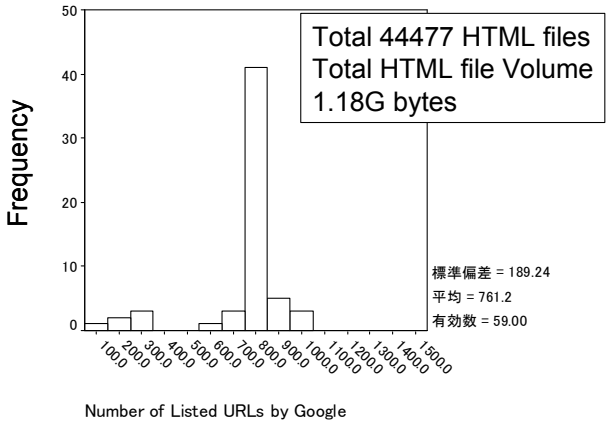

Fig. 11: Frequencies of Valid URL List for 59 Cancers

を含んでいることが推測された．以下、これら単語集合それぞれの重複語彙数を Table 3 にまとめた．

3.5 対象とするがんＵＲＬ集合と解析用データの固定

Table 2 に挙げた 59 疾患それぞれを検索語として、検索エンジン(Google!)を用いて URL リスト(それぞれの疾患において約 800－900 個程度)を取得、合計 44477 個の Valid な URL リストを得た．本 URL リストをもとに、wget を用いて、同数の URL ツリーを全量ダウンロードし、HTML ファイルとして合計 1.18 ギガバイトを得た．概要を Fig.8 に示した．本データを対象として、G, M, C1, C2 それぞれの重複語集合の出現頻度と出現 URL 数をカウントした．結果を Table 4 に示す．

3.6 各用語集合の特性に関する検討

3.6.1 出現頻度

各語に関して 1 回以上出現したものを”Appear”とし、重複語数との比を Appearance Rate とした．結果を Table 5 に示した．Table 6 から本 URL 集合において、G に比べて M,C1,C2 の Appearance Rate（出現比率）が

Table 5: Various Parameters of 59 Cancer-44477 URLs for Each Word Sets (G, M, C1, C2 and Various Duplicate Data Sets)

Category of Sets	N of Words (A)	N of Appearance (B)	Appearance Rate (C)	Total Numbers of Frequency (D)	Total Numbers of Appeared URLs (E)	Mean Frequency / word (D/A)	Mean Appeared URLs / words (E/A)	Measured Mean Frequency / word (D/B)	Measured Mean Appeared URLs / words
C1	3315	2470	0.745	315469	117107	95.16	35.33	127.72	47.41
C2	9451	5826	0.616	1356535	407453	143.53	43.11	232.84	69.94
M	59533	40381	0.678	3954519	1398453	66.43	23.49	97.93	34.63
G	75498	43942	0.582	11436557	2517477	151.48	33.34	260.26	57.29
C1∩C2	2494	1948	0.781	231973	92548	93.01	37.11	119.08	47.51
C1∩M	801	762	0.951	171140	64655	213.66	80.72	224.59	84.85
C1∩G	103	101	0.981	26599	9658	258.24	93.77	263.36	95.62
C2∩M	1974	1870	0.947	804554	235607	407.58	119.36	430.24	125.99
C2∩G	1046	1021	0.976	665735	157076	636.46	150.17	652.04	153.85
M∩G	2827	2588	0.915	816454	236606	288.81	83.70	315.48	91.42
C1∩C2∩M	686	657	0.958	157981	59483	230.29	86.71	240.46	90.54
C1∩C2∩G	66	64	0.970	21370	7391	323.79	111.98	333.91	115.48
C1∩M∩G	69	69	1.000	22966	8068	332.84	116.93	332.84	116.93
C2∩M∩G	629	626	0.995	465793	110956	740.53	176.40	744.08	177.25
C1∩C2∩M∩G	48	48	1.000	19809	6821	412.69	142.10	412.69	142.10

高いこと、中でも G の 0.58 に対して C1 の 0.74 は語彙数が G に比べて 20 分の 1 程度であるにもかかわらず高い事から本 URL データにおいて特異性が高いことが推定された.

また、C1∩M、C2∩M が、それぞれ語彙数 801 と 1974 に対して C1∩C2∩M が 686 と C1∩M の率が高かった. また、C1∩C2∩M と C2∩M∩G の語彙数が両方とも 650 前後であり、C1∩C2∩G と C1∩M∩G がほぼ同数の 66－69、出現数もほぼ同程度であったことから C1∩M が特異的に出現していることも推定された.

3.6.2 C1, G を構成する語に関する検討

前節での検討により、C1 が語数の少ないこと、特異度が高いことからがん情報コンテンツの内容把握や評価を与える目的において有用であると思われた. そこで、C1 を構成する語彙の特徴を一般用語集合 G と比較した. C1 と G の単語長と対象とする 44,477URL における出現頻度を Fig. 9 に、それぞれの特徴の概要を Fig10 に示す.

Fig9 より, C1 は G に比べ、5 文字から 10 文字の中等度に長い単語において、出現回数が 1000 をこえるものがあることがわかる. C1 においてこの単語長では、骨髄性白血病, 悪性リンパ腫, HTLV-I, CA19-9, 多発性骨髄腫, 腫瘍マーカー, CPT-11 などが出現しているのに対して、G ではアレルギー, クリニック, ランキング, 問い合わせ, ダイエット, アスベスト, ありがとう, アルコール, カテゴリーなどのカタカナ語であり、両者の語彙集合の質は全く異なることがわかる. さらに 10 文字以上の語において G ではコミュニケーション, リハビリテーション, インフォメーション, インターナショナル, インキュベーション, エンターテイメント, プロフェッショナル, ワークステーションなど全てカタカナ語であったのに対して、C1 では非ホジキンリンパ腫, 急性リンパ性白血病, セカンドオピニオン, 慢性リンパ性白血病, シェーグレン症候群, パピローマウイルス, モノクローナル抗体といった、がん情報に関連した語であった. 上記の G に含まれるカタカナ用語と C1 に含まれる語は意味内容と医学的知識への適合性が異なっており、がん用語辞書 C1 を用いた方がより生理的に理解しやすいと思われた.

この結果は、統計上はがん専門用語集合 C1 は、本研究の最初に問題として認識した、がん URL 群の解析において一般用語辞書を用いた場合でもがん専門用語辞書を用いた場合でも、用いた単語集合の検出力によっては特にベイズ分類器などの統計手法を用いた分類において特徴量を計算し学習する事が可能だが、その分類（例えば分類スコア、ベイズ分類器の各集合への存在確率）の内容は目的としたコンテンツ内容による分類（人間が意味内容に基づいて分類している分類）とは異なる

3.6.2 各用語集合の検出力に関する検討

得られた単語集合(C1, C2, M, G および重複各集合)それぞれの総合的な検出力を比較するため、要素の出現頻度(D)と対象 URL 群(がん URL 群、44,777 URL)の中での出現 URL 数(E)の各集合での総和、分母を語集合の語数とした総出現頻度率と出現 URL 率（D/A, E/A)、出現した語数のみを分母とした場合(D/B, E/B)を表 6 に示した. 例えば D/A はその語集合を構成する単語の対象 URL 群における平均出現率である. G,C2,M,C1 の順であり、2 群の重複集合（例えば C1∩C2, C2∩M など）では C2∩M,C2∩G の方が M∩G よりも D/A 値が高い. さらに 3 群の重複集合では C2∩M∩G の D/A 値が 740 と他の場合が 300 前後であるのに比べ 2 倍以上であった. 出現した単語数を分母とした(D/B, E/B 値)においても同様の傾向がみられた. これらのことから、解析に用いる用語集合の検出力は C1 に比べ C2 の方が高いことが示唆された.

以上のことから、語集合 C1 によりがん情報 URL は一般用語辞書を用いた場合に比べて的確に内容を把握できることが示された. 今回の C1,C2,M,G の言葉集合をもとに、コンテンツ評価用ワードセットの作成に関して検討することとしたい.

まとめ

がん情報提供状態の質的評価を定量的計測を可能にする外的基準を与えるべく、がん専門用語集合について検討した. がん用語集合は一般に認められ広く用いられている国立がんセンター(NCC-CIS)の各種がん

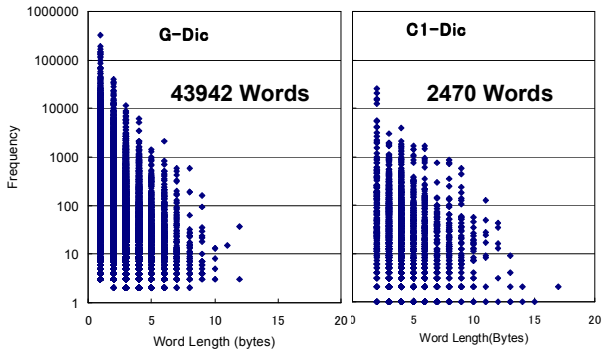

Fig.9: Comparison between G and C1 by Frequency of 59 Cancer (44,477 URLs)

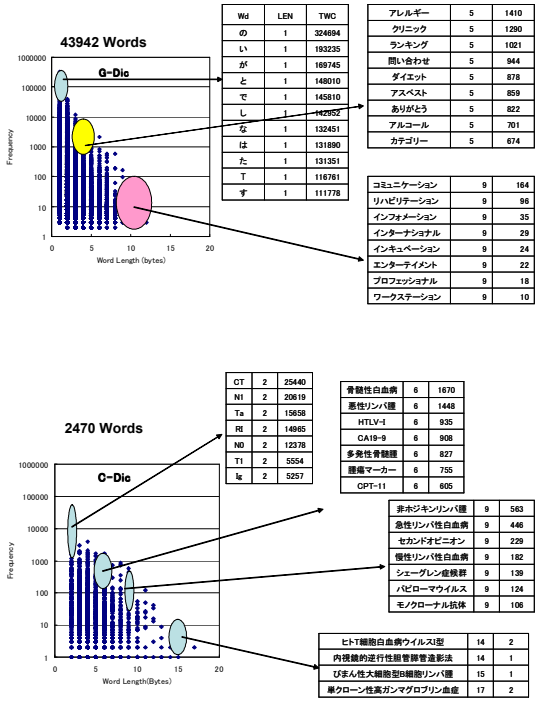

Fig.10: Detailed Comparison between G and C1 by Frequency of 59 Cancer (44,477 URLs)

（54 種）の診断、治療ならびに説明を行なっているコンテンツから手で切り出して作成した（C1）．その結果、C1 の語数は約 3400 語であり、一般用語辞書の約 7 万、医学用語辞書 M の約 6 万に比べて小さかった．がんセンターの提供するコンテンツ改訂に従い約 3 万語の抽出を行い、約 9600 語の C2 を得た．

C1,C2 の特性を検討するために、各種がんを検索後として既存の検索エンジンで得られる URL リスト（約 45000 個）を元にデータを全量ダウンロードし、HTML のテキストデータとして約 1 ギガバイトのデータを得た。このデータに対して C1,C2,M,G の 4 つの言葉集合それぞれの語の生起頻度を計測し、出現率で比較したところ、C1 は一般用語辞書 G の 58%に比べて約 75%と高く、本集合ががん情報コンテンツの内容に特異性が高い事が示された．

以上のことから、今回作成した C1 は、G、M に比べて語数が約 10 分の 1 であり、評価計算の負荷を大幅に減らす事が可能でありがん情報コンテンツへの特異性が高い事から有用である事が確かめられた．

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（新世代ネットワーク研究センター），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] 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

[3] 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

[4] 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

[5] 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

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

[7] 国立がんセンター http://www.ncc.go.jp