介入的手法によるがん情報取得適正化に関する検討

中川晋一[†‡*]、木村俊也[‡]、三角真[*]、島津明[‡]、山岡克式[†]、酒井善則[†]

† 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
‡ 北陸先端科学技術大学院大学 〒923-1292 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

E-mail: † {nakagawa, yamaoka, ys}@net.ss.titech.ac.jp, ‡ {s-kimura, shimazu}@jaist.ac.jp, *{snakagaw, misumi}@nict.go.jp

あらまし がん患者にとって Web を介した情報は重要である。調査により、欧米に比べわが国では、コンテンツ量で 100 倍の違いがある上に患者への配慮の点で遅れている、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であることが特徴であった。患者の得る情報の充実のために個人による情報発信を用いることが有用であるが、一般の検索エンジンによって提供される URL の順位は商用サイトへの誘導などを含んでおり適切ではない。適切な URL リストを提供するための介入的手法のフレームワークについて検討した。(1) 形態素解析による頻出語順の言葉空間、(2) 表層からの計測項目（該当 URL の第一層の URL データ構造：全データ量、HTML ファイルデータ量ならびにイメージデータ量とそれぞれ全体との比など）を説明変数として適切なスコアリングを与える手法(Fairly Index)を与えるための必要用件について検討した。

キーワード がん情報、Web データマイニング、Web セマンティックス

1. はじめに

Web を介する各種のがんに関する情報提供は重要であり、多くの医療従事者や患者に対して行われている。がん情報が他の医療情報に比べて盛んに流通するのは、未だに不治の病であることが原因である。がんを宣告された患者や家族は少しでも新しく、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る必要から情報検索に対する要求が高い。理由として治療法・診断法の確立されているほかの疾患に比べ、昨今報道されているように患者自身が正確な情報を知る権利が注目されており、どこに行けば標準的な治療がうけられるのか、自分の聞いている病態は本当に正しいのかなど、さまざまな情報要求が高まっている。

しかし、Yahoo!等検索エンジンで「胃がん」の検索に対して数百万のヒットがあるが、上位 100 位に出現する URL は必ずしも患者の要求にかなうものではない。特に商用サイトへの誘導、通信販売による健康食品への誘導は深刻である。また、Web 上で従来行われてきているこの分野の情報提供は医-医間の情報交換が目的であるものが少なくなく、難解な専門用語が多く理解するために専門知識を要する。患者が大学・研究機関のページにアクセスしても必要な情報を得られる可能性は低い。

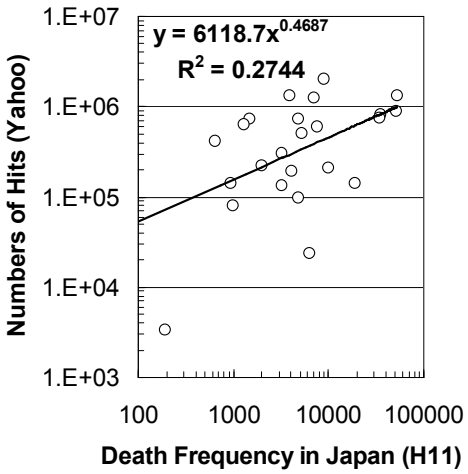

Fig.1: Various Cancer Death (H11) and Searched Number of Hits

医学分野での Web データマイニングはベイズ法を用いた分析[1]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 Web 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[2]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。

本研究では、患者のためのがん情報の適切な提供を目的とし、既存のサーチエンジンで得られた URL を中立的なフレームワークでスコアリングしなおすことによって、より適切な情報提供を行うことができる必要条件を検討する。

1.1. わが国におけるがん情報流通状態の特徴

わが国におけるがん情報提供の状態の一例を Fig. 1 に示した。これは、平成 11 年における国際疾病分類（ICD-10[3],[4]）による部位別の死亡数（1 年間の各項目についての死亡数：総数約 29 万人）と Yahoo を用いてそれぞれの項目に該当するがん（例えば、ICD で胃であれば、「胃がん」）で検索した場合のヒット数を示す。シンボル一つががんの種類 1 つに相当する。$R^2=0.27$, $p<0.03$ で有意な相関関係を示した。また、

年間の死亡数が 1000 件未満のものであっても 3000 以上のヒット数が観測された。この分野に関する情報提供が活発である事を示している。

1.2. 合衆国との比較

インターネットでの情報発信がわが国よりも先に始まった合衆国とわが国のがん情報コンテンツ提供の状態を比較した。

1.2.1. コンテンツ量

各がん専門機関で提供されるデータコンテンツ量の日米比較を行った。対象としたのは、全国がん協議会に加盟する 30 のがん専門医療機関と全米トップ５機関(NCI, Sloan - Kettaling C.C., M.D. Anderson C.C など)である。結果を Table 1 に示す。Web データの収集には LinuxOS で wget を用い、再帰的なダウンロードをしない条件でそれぞれの機関の URL のトップページから行った。総コンテンツ量で２桁の違いがあるのに対して総イメージファイルデータ量は約３倍と、コンテンツ総量の違いが HTML ファイルそのもののデータ量（文字数とファイル数）によることが示唆される。また、NCC（国立がんセンター）と NCI（National Cancer Institute）で提供されている疾患項目がそれぞれ 50 対 201 であった。

1.2.2. コンテンツ提供形態の日米差

これら日米差の原因を探索する事を目的として、2005 年 8 月一般に用いられる Web 検索エンジンを用い、NCC と NCI で共通に情報提供が行われている傷病名について（胆管がんと bile cancer を用いた）得られた URL リスト 100 個を分類した。結果を Table 2 に示す。US では Cancer Net などの authorized institutions が Peer Review を行っていると思われるコンテンツ提供が全体の約半数を占めるのに対して、わが国では Peer Review されていない病院や個人のコンテンツが多い。わが国でのがん情報提供の特徴として、Peer Review を行っているような情報は少ないが、むしろ個人的に行われている情報提供が多い。特に患者コミュニティを形成するために重要な個人の闘病日記のようなコンテンツが数多く見られることが特徴である事が示唆された。US において Peer Review を行った情報提供が約半数を占める理由として、インターネット情報提供がすでに常識化しておりシステム化されている（専門部署が存在する）ことなどが考えられるが、個人による情報発信（例えば闘病記の類）は殆ど発見する事ができなかった。

	Japan mean ± S.D	United States Mean ± S.D.
Total Volume of Contents(MBytes)	0.4 ± 0.5	1.2 ± 0.9
Number of Files(x1000)	1.2 ± 1.5	36.7 ± 31.7
Number of HTML Files(x1000)	0.4 ± 0.6	34.2 ± 32.3
Volume of HTML Files(MBytes)	4.3 ± 7.3	948 ± 844
Number of Still Image Files(x1000)	0.7 ± 0.8	2.1 ± 1.9
Volume of Still Image Files(MBytes)	12.5 ± 15.3	28.3 ± 28.5
Number of cases	30	5

Table1: Comparison of Cancer-Web Mining Data between 30 Japanese Special Cancer Facilities and 5 US Facilities at 2005

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

Table 2: Proportion of Contents Distributors for Bile Cancer between US and Japan at Top 100 Hits at 2005 August.

2. ユーザ指向の URL リスト適正化手法

2.1. 従来手法

Fig.2 に従来行われてきたキーワードサーチを行う機構を示す。初期は、ロボットにより、殆ど全ての公開されている Web サーバから HTML ファイルをダウンロードしてデータベース化し、ユーザの要求するキーワードによってサーチ、ヒットした URL リストを提示する仕組みであった。この場合、User からのアクセスの高いキーワード、クリックされた頻度の高い URL がリスト内で繰り上がる。このため、作為的にある URL ばかりをクリックするようなバイアスが人為的に加わった場合は、目的とするリストに狂いが生じる。本法は原則として対象とする URL ツリー全てをダウンロードする必要があるため、膨大なディスク領域を

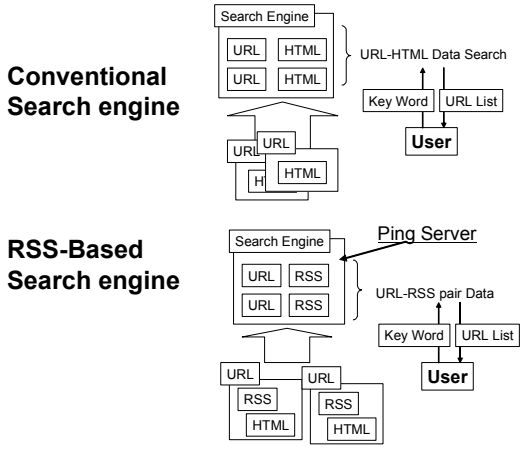

Fig.2: Models of Search Engine by Keyword; Conventional type and RSS based.

必要とする。また、原則的にフルテキストの逐次探索であり、サーバ側の負荷も大きい。RSS と呼ばれる XML で記述可能なメタデータ記述様式が注目されている。本法は、あらかじめサーチされたいキーワードを XML に書いておき、Ping サーバと呼ばれる URL-RSS ペアを保持するサーチエンジンのデータベースに登録しておく。ユーザの要求したキーワードを RSS データベースの中から検索し、該当する URL リストを返す。RSS はキーワードであり、HTML のフルテキストに比べ、記憶領域は小さい。また、キーワードの設定は発信側の情報コントロール権に立脚したモデルである。RSS キーワードデータベースの形態でデータが保存されているため、検索時に複雑な HTML の構造解析を必要とせず、サーバの動作が軽いのが特徴である。しかし、既存の RSS はキーワード設定のアルゴリズムを公開すると同時にターゲットになることや、悪意を持った管理者が作為的に大きな Key Index を記述するなどの問題に対しては解にはなりえない。

2.2. Fairly Index Server の提案

今回の調査によって、既存方式で与えられるリストは作為的操作によって本来の目的から外れることが問題である事が示唆された。簡単なキーワード選択には限界があり、(1) Positive Word と Negative Word、(2) 改変可能な禁止論理、(3) ユーザの指向に合わせたフィルタリングの 3 点を実現する必要がある。

このため、RSS のような自己申告型ではなく、外部監査型システムで調整論理をユーザ側の設定に従って変化しうるシステムが求められる。しかし、Search engine のようにいつ来るとも分からないユーザのために大規模なディスク空間に常に逐語検索のための HTML ファイルを保持するのは困難である。

以上のことから、介入のポイントとして、ユーザがサーチエンジンから URL リストを受け取った時点でキャッシュサーバから問い合わせ、URL リストの中でそのユーザがあらかじめ設定した「レベル」に応じてユーザへの回覧を許すかどうかという選択を行うことが考えられる。例えば、レベル１：学術情報だけを得る。２：個人的な情報発信データを得る。３：ポータルサイトや書籍など情報のためのデータを得る。４：商用サイトまでのデータを得る。５：全てのデータを得る。という動作が必要である。

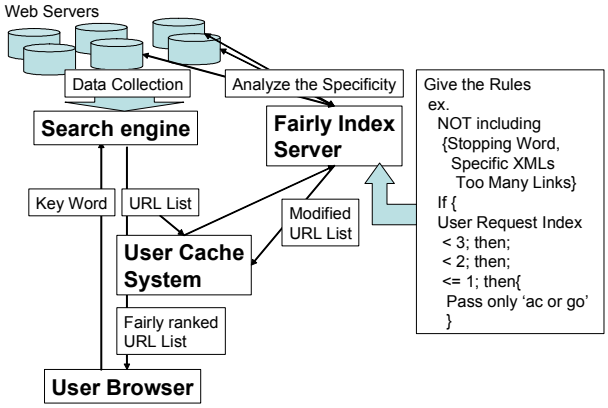

Fig. 3: Overview of 'Fairly Index Server'

レベル１の学術情報はわが国では、ac ドメイン、go ドメイン、pref ドメイン、一部の or ドメインから発信されており、co ドメインから発信される事はない。しかし、日米比較からこの種の情報は量が少なく専門的であり、実際に患者コミュニティ形成などのニーズには応じられるものではない。むしろ医師個人、病院、患者や家族による闘病記の方がニーズが高い可能性が高い。従って、この例のレベル２、３を４，５の営利目的のものから選別する必要がある。これらは単純なキーワードの条件設定、ドメイン分析では実現する事が困難である。実際に与えられた URL をキーワードに対して何らかの適合性スコアを算出しレベルに応じてユーザに選択させるというフレームワークが望ましい。以上のことからユーザのブラウザの通信に介入し、サーチエンジンから得た URL リストを並べ替える Fairly Index System(Fig. 3)を提案する。

ユーザへの通過ロジックは Fairly Index Server 側にあるため、状況の変化にも対応しやすい等の利点がある。また、従来型のサーチエンジンとは異なり、本サーバはユーザによる逐次の URL リストに対して一定のアルゴリズムでユーザの設定したアルゴリズムに基づいて URL の順番や参照の可否に関して介入するため、HTML コンテンツそのものの検索の機能は必要ない。例えば、" Neglect Unreachable : Unreachable な URL を除く"という設定を行った場合に Ping を送出して一定時間内に echo reply が帰らなければその URL を削除して返すというような動作を想定する。

3. Fairly Index 作成のための検討

Fairly Index Server は、問い合わせられた URL が、どのような特性をもつのかを出来るだけ適切にスコアリングすることを目的とする。実現のために、(1)現行のサーチエンジンによる検索結果の状態、(2)内容の教師データを与えるためのスコアリング、(3)特徴づけのための言葉集合の特定、(4)対象とする URL の形態（トップページのデータ量、URL に含まれる HTML ファイルや画像ファイルの大きさなどの形状分析結果）の各項目について検討した。これら項目をどのように用いて Index を作成すれば良いかの方向性を探索することとした。なお、「がん」は病態の異なる 30 以上の疾患

群（合衆国の NCI の提供している疾患名としては 200 以上）であるが、今回は大腸がん（CC: Colon Cancer）、胃がん（SC: Stomach Cancer）、肺がん(LC: Lung Cancer)、子宮がん(Uterine Cancer)、白血病(Leukemia)の５つのがんを対象として検討した。

3.1. 調査と方法

通常用いられる検索エンジンを用い、得られた URL リストを固定、URL リストの示す HTML ファイルをダウンロードし、対象とする HTML ファイルを固定する。これらに対して評価者がカテゴリー分けを行いそれぞれの URL のスコアを決める。このスコア(教師データ)に対して単語分析、URL データの性状分析等によって得られた各種パラメータとの関係を検討することによって、Fairly Index を一定のアルゴリズムで分類できる方式を検討した。

3.1.1. URL リストの固定

検索エンジンとして www.yahoo.co.jp を用い、キーワードとして(1)胃がん、胃ガン、胃癌、(2)肺がん、肺ガン、肺癌、(3)子宮がん、子宮ガン、子宮癌、(4)大腸がん、大腸ガン、大腸癌、(5)白血病をそれぞれ OR 条件で入力し、それぞれの傷病名に対して 1000 個の URL リストを得た。

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 位までにランクされたものを対象として、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。

3.1.3. 分類（教師データの作成）

得られた HTML ファイルを一つ一つ、医師（専門的知識を持つ）、がん患者(ある程度専門的知識をもつ)、学生（専門的知識を持たない）の３名で順不同別々に次の分類を行った。C-1 から C-5 まで、それぞれのカテゴリーについて CII (Cancer Information Index)の１から５とした。

C-1: Peer Review を行っていると思われるがん専門機関による情報；がんセンターや大学病院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 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、いわゆる blog 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；ヘッダやフッタに目的とする用語が含まれたりしていて得られた HTML ファイルには検索語が見つからないもの。

本スコアリングにより CII は増加するほど大きくなればなるほど専門的ではない、検索ノイズへと変動する。また、Table 2 の検討により、わが国の特徴として専門的情報発信を行っている C-1 カテゴリは少なく、C-2 が医師個人ならびに患者個人であり発信者の母集団が最も多いことが予想され、医師個人によるものと患者個人によるものを分別するためのサブカテゴリを設定する事も検討したが、今回はフレームワークの検討を目的としたため、クラス別の度数に関しては考慮しないものとした。

3.1.4. 専門語辞書の作成

それぞれのキーワード（胃がん、肺がん、大腸がん、子宮がん、および白血病）の傷病名に対して、国立がんセンター（NCC-CIS）で提供されている HTML からテキストデータを得た。これから、普通名詞、形容詞（明るい等の通用される形容詞）、動詞、副詞を除き、医学用語として用いられる名詞、形容詞（侵襲的など）の用語を切出し辞書作成した。なお、子宮がんの場合、子宮頚部がんと子宮体部がんの OR データ、白血病の場合、急性骨髄性白血病、慢性骨髄性白血病、急性リンパ球性白血病、ホジキン病、悪性リンパ腫など疾患概念が複数になるものがあるが、白血病に関しては急性骨髄性白血病と慢性骨髄性白血病の２つの OR データを用いた。

	Total Volume of Files(Bytes)		Number of Total files	
	Mean	S.D.	Mean	S.D.
Lung Cancer	1.1E+05	1.5E+05	8.5	11.8
Leukemia	8.7E+04	2.1E+05	11.3	21.6
Colon Cancer	8.7E+04	2.1E+05	12.4	13.6
Stomach Cancer	7.4E+04	8.4E+04	10.3	11.3
Uterine Cancer	7.3E+04	8.1E+04	11.1	13.4

Table 3: Web Data Contents Volume of 100 URLs of Each Cancer Keyword in Japan.

Category	LC	Leu	CC	SC	UC	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

Table 4: Result of URL Categorization for Five Cancers.

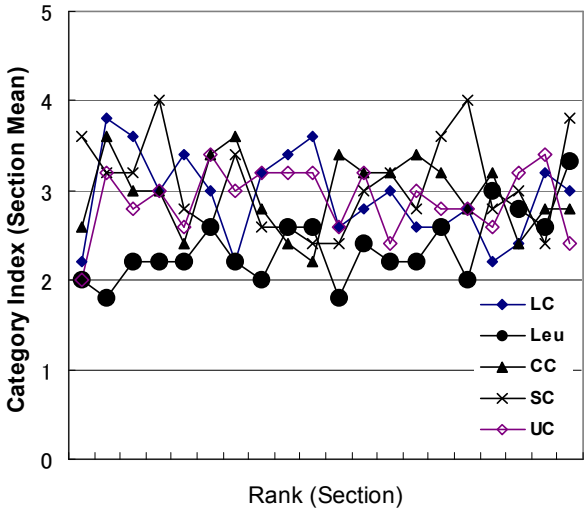

Fig. 4: Section Mean (/5URLs) of Category Index by Listed Rank of URLs

4. 結果と考察

4.1. HTML ファイルの取得結果

前述の方法によって Yahoo.jp で得られた URL（合計 1000 リスト）を元に、必要箇所を抜き出し、それぞれの URL に対して Wget を用いて、各 HTML ツリーをダウンロード、それぞれの傷病名別の総ファイル容量（Total Volume of Contents）と総ファイル数を Table 3 に示した。傷病名別では肺がんが多く、全がんのトップ URL での取得データ量の平均値は 110 Kbytes ファイル数は平均 10 個であった。

4.2. URL 分類の結果

Table 4 に傷病名別のカテゴリー分類の結果を示す。C-1 5%、C-2 43%、C-3 21%、C-4 24%、C-5 は 7%であった。C-2 は個人闘病記と医師を含む個人による情報発信を含むため、高率となった。C-1 と C-2 の和と C-3、C-4 の和はほぼ等しかった。

Fig.4 にそれぞれの傷病での 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 の順位毎の 5URL ずつを区切りとした区間のスコアの平均値の変動を示した。傷病別に特徴が見られ、特に大腸がん、胃がんでは上位ほどスコアが高く、白血病では順位に従ってノイズが増加した。これは Fig.1 で示したように、それぞれのがんによって死亡数（発生数）が異なり、その発生数によりヒット数が多いことから、「一般的な」胃がんや大腸がんに比べ、発生頻度が少ない白血病がターゲットキーワードとして未だ使われていない可能性が示された。

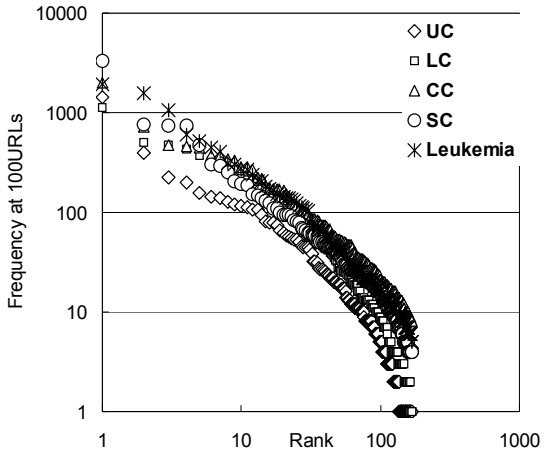

Fig. 5: Frequency of Top 128 Words by Each Cancer Categories

4.3. NCC-CIS 言葉集合の特性

がんの種類は多岐に渡るが、「診断方法（レントゲン写真、CT 検査、MRI）、治療方法（外科的切除、化学療法、緩和ケア）」等の語は共通する。また、病理学的（細胞レベル）では扁平上皮がんであっても、出来た場所によって肺がん、喉頭がん、咽頭がんに分類され、治療方法も外科切除、化学療法、放射線である。従って、これらの疾患に関して説明する場合、項目が別であっても診断法や治療法が同じ（説明の中で用いられる単語が共通）であることが多い。これらの言葉遣いは教育を受けた専門家が書いたものとそうでないものでは明らかに異なる。解析に使用する標準語群として、わが国で標準的に使われている NCC-CIS で提供されている Web ページをそれぞれのがんについて取得、単語を切り出した。その結果、Leukemia、LC, SC,

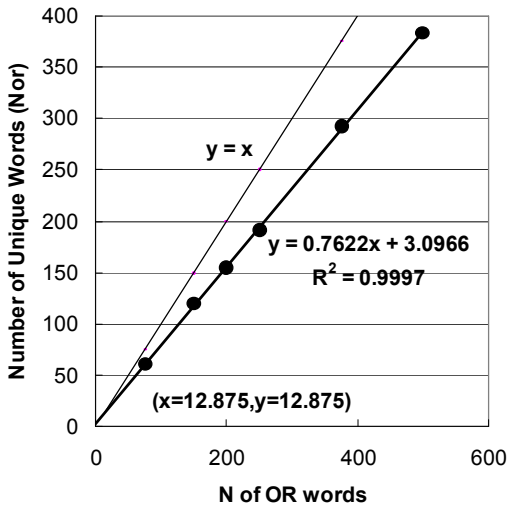

Fig. 6: Result of ‘Unique’ for the simple ‘or’ of Various Rank of 5 Cancers

CC, UC の５種類の各がんについての説明文書に出現する単語はどれも約 300 語であった。これら語それぞれを対応するがんを検索語として得られた URL 毎の出現率で順位付けした。それぞれのがんにおける上位 128 個の出現頻度を Fig.5 に示す。どのがんについても出現頻度は上位 20 個までで 100 以下となった（1 URL あたり 1 回)。この事から出現する単語自体は 1 がんあたり 20 個以下であることが示唆された。さらに、それぞれのがんの上位 15(５つのがんで合計 75 個), 30(150), 40(200), 50(250), 75(375), 100(500)に出現したワードの Or を取ったものを x 軸（重複を含む）とし、それらの言葉群の単語の異なり数を y 軸としてプロットした。結果を Fig.6 に示す。NCC-CIS はがんの専門家が作成し、Reviewer が一般人が読みやすいように改変されている。そのため各種がん別に使用されている用語が統一されており、がんの個数に関わらず出現する特殊語の個数も一定する。また、本コンテンツに用いられている用語がそれぞれの検索語によって選択される URL において均一に用いられていると考えた。これらから、それぞれの疾患別に出現する単語の中で頻度の高い順に同数をとり、評価用語群とした。Fig.6 から y=x と実測値の交点を求めると約 12 となったため、各疾患別に上位 100 個の URL での出現頻度上位 12 個の単語をそれぞれ選択し計 60 個として(それでも手術という単語が重なったため使用したのは５９個）を用いて言葉集合を決めた(Table 5)。

4.4. 単語出現頻度（WFI）による分析結果

Fig.7 に 1 から 59 までの番号と CC, UC, LC, SC,

No.	Word	No.	Word	No.	Word
1	ポリープ	21	急性骨髄性白血病	41	結腸
2	リンパ性白血病	22	手術	42	肛門
3	上皮	23	抗がん剤	43	肝
4	乳がん	24	放射線	44	肝臓
5	内視鏡	25	染色体	45	肺
6	再発	26	検査	46	胃
7	出血	27	検診	47	胃がん
8	切除*	28	治療	48	脳
9	前立腺	29	潰瘍	49	芽球
10	副作用	30	症候	50	血小板
11	化学療法	31	療法	51	血液
12	卵巣	32	白血球	52	診断
13	喫煙	33	白血病	53	転移
14	大腸	34	白血病細胞	54	進行
15	大腸がん	35	直腸	55	頸部
16	婦人科	36	神経	56	食道
17	子宮	37	移植	57	骨
18	子宮がん	38	粘膜	58	骨髄
19	子宮内膜	39	細胞診	59	骨髄性白血病
20	寛解	40	組織		*:「切除」のみ重複

Table 5: Selected 60 words for WFI (Word Frequency Index) at Five Cancers by Top 15 frequency of 100 URL lists of each Cancer Category (Stomach Cancer, Lung Cancer, Uterine Cancer, Colon Cancer and Leukemia)

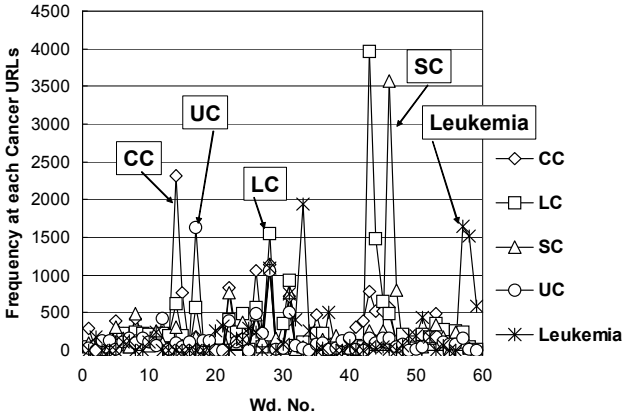

Fig.7 Selected 59 words spectrum for each cancer 100 URLs

Leukemia それぞれの検索結果のうち 100 位までの URL における出現頻度を示した。特異的に頻度の高い単語があり、それぞれのがんで重なっていないことが分かる。ピークを示す「白血病」などの言葉は 100URL あたり 1500 回であり、1 URL あたり平均約 15 回の出現

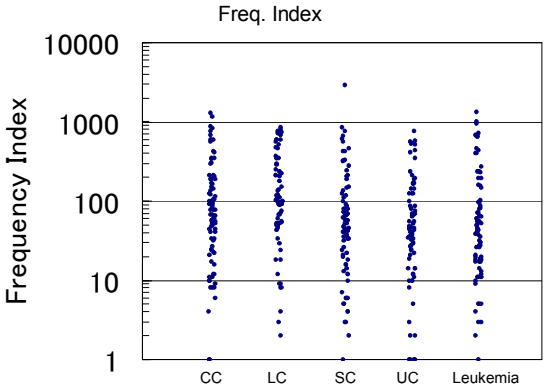

Fig. 8: WFI by various Cancer 100 URLs.

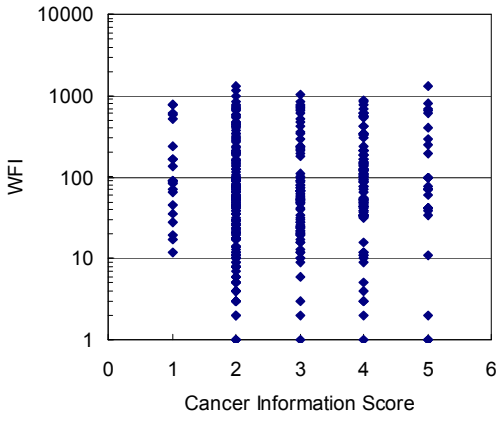

Fig.9: Plot of CII (Cancer Information Index) and WFI (Word Frequency Index : SUM of five cancers)

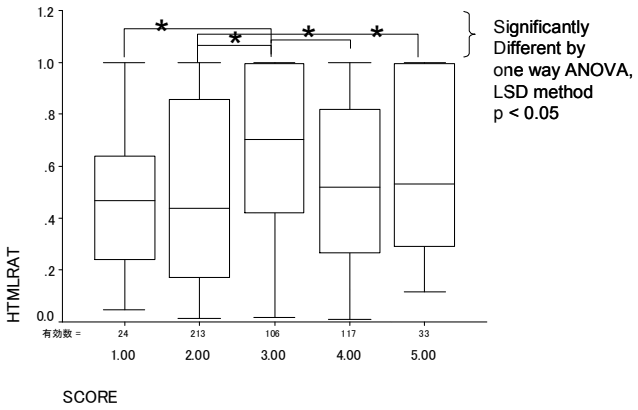

Fig. 10: Result of One Way ANOVA for CII and HTML rate

頻度である事を示している。今回行った単語集合の作成方法により、それ程多い単語数ではなくてもそれぞれのがんの特徴を記述できることが示された。また、本言葉集合の出現頻度の合計値を WFI(Word frequency Index)とした。それぞれのがん別の本値を Fig.8 に示す。また、URL の教師スコア（CII: Cancer Information Score）に対する WFI のプロットを Fig.9 に示す。CII を目的変数とした 1 元配置分散分析の結果、Sheffe の多重比較法における有意性確率は p=0.058 であり統計学的有意性は検出できなかった。

4.5. Web コンテンツ計測結果と CII の関係

さらに、付加的に Web コンテンツの構成にも着目し、パラメータを検討した。対象とした 5 つのがんについてそれぞれ 100 ずつ計 500 のページを対象として、Web データの形態的計測結果 7 値（総データ量 、総ファイル数、総 HTML ファイル数、総 HTML ファイルデータ量、総イメージファイル数、総イメージデータ量、および HTML 比：HTML データ量/総データ量）に対して CII を目的変数として同様に 1 元配置分散分析を行った。その結果、Sheffe による多重比較で HTML 比が CII に対して有意(p<0.05)であり、1-3, 2-3, 2-4, 3-5 間で有意差が認められた。これらのことから、HTML 比が CII=2：個人による情報発信と他のものを分別するために有効である事が示された。

4.6. Fairly Index 算出の検討

以上の検討からがん情報を提供しているページの質の中で強調したい CII=2 の分類のために必要な要件として、WFI だけではなく HTML rate を選択することが有効である事が示された。WFI の適正化は今後の検討課題である。CII を目的変数として NCC-CIS のページをもとにベイズ法を適応し分類も行ったが、この場合は教師データに対して良好な結果（正答率 80%以

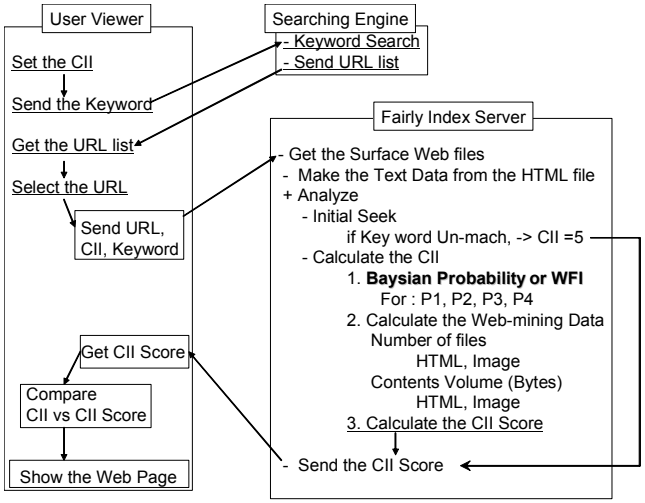

Fig.11: Summary of Processing Flow calculating the CII Score

上）を得たことから、本言葉空間を用いた CII の分類は有効であると思われる。ベイズ法は処理のオーバヘッドが大きいこと、本検討から明らかになったように各がんの特徴語は高々20 語程度であることから、単語数を適正化する必要はあるが、最終的には本単語群を用いた決定木的な手法の方が処理の軽いことから実装に適している。ベイズ法の適応ならびに変更に関しては木村[6]の報告で検討した。ベイズ法での CII=2 の選択に関して教師データに対する正答率は 80%をこえたことから実用レベルにあると思われる。しかし、本法では Fig.11 に示すように、ユーザがひとつひとつの URL を選択しようとする時に URL、CII、Keyword を送付し、サーバ側で CII を計算することを想定するため、処理時間が問題であり、できるだけ処理の軽い方式を検討し実装するのが今後の検討課題である。

5. まとめ

Web を介したがん患者のための情報提供状態の改善を目的として以下の検討を行った。

1．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少数であり、コンテンツ量も合衆国に比べて大幅に少ないが、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であることが特徴であった。

2．胃がん、大腸がん、肺がん、子宮がんおよび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分類しスコアリング(Cancer Information Index)した。その結果、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おり、疾患の発症率の低い疾患ほどノイズが低く、高い

疾患ほどノイズが高く検索順位に対するバイアスがかかっている可能性が示唆された。

3．これらのことから、これら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。特に個人レベルでの情報発信の質を自然言語学的に分類する手法について検討した。

4．標準的な単語集合の作成を目的として、わが国で標準的な情報提供を行っている NCC-CIS のコンテンツをもとに単語のリストアップを行った。分析の結果､疾患別に特異的な単語の語数は 12 語程度ではないかと推定した。本単語集合の有効性の検証を目的に、対象としたそれぞれのがん 100 個の URL におけるそれぞれの単語の出現頻度の合計スコアを WFI(Word frequency Index)として算出し、CII との相関を検討したが統計学的有意性は検出されず、今後の検討課題とした。

5．Web 提供状態の計測結果からダウンロードデータ中の HTML ファイルのデータ量に対する比（HTML 比）が、CII に関して分散分析・多重比較の結果有意であり、本指数が分類に関して有効である事が示唆された。

6．これら言葉集合を用いたベイズ法による解析の結果､高い正答率が得られ WFI のさらなる適正化によって実行時の負荷が小さい（適正語集合に対する単語出現頻度でも示唆可能な）システム構築の可能性も示唆された。

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] 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

[2] http://www.jima.or.jp/

[3] F.Fukuda, Y. Ohashi, "A Guideline for Reporting Results of Statistical Analysis inJapanese Journal of Clinical Oncology", Jpn J. Clinical Oncology, 27(3), pp121-127, 1997

[4] The Research Group for Population-based Cancer Registration in Japan, "Cancer Incidence and Incidence Rates in Japan in 1998: Estimates Based on Data from 12 Population-based Cancer Registries, Jpn J Clin Oncol 2003;33(5)241-245, 2003

[5] Jinqiu Guo,Akira Takada,Koji Tanaka,Junzo Sato and et.al,"CLAIM (CLinical Accounting InforMation) An XML-Based Data Exchange Standard for Connecting Electronic Medical Record Systems to Patient Accounting Systems", Journal of Medical Systems, Vol. 29, No. 4, pp413-423,2005

[6] 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-i9, 2006