患者のためのがん情報 URL リスト適正化に関する検討

Survey Results of Reasonable URL List Providing for Cancer Patients in Japan

中川 晋一♥ 木村 俊也♦ 三角 真♥
島津 明♦ 山岡 克式♠ 酒井 善則♠

Shin-ichi NAKAGAWA Shunya KIMURA
Makoto MISUMI Akira SHIMAZU
Katsunori YAMAOKA Yoshinori SAKAI

がん患者にとって Web を介した情報は重要な社会情報基盤となりつつある。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。本研究では、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を与えるための因子（1）形態素解析による頻出語順の言葉空間、（2）表層からの Web マイニング計測データ項目などに関して検討した。

The importance of Web-based Cancer information is becoming as the basic social infrastructure for the cancer patients and their families. Because of the cancer is one of the still incurable diseases for the human race, more progressive information will give the better possibilities for survivals. The comparing survey between US top five cancer facilities and Japan carried out that the volume of Japanese were less than 1/10. And another survey of comparing top 100 URLs between those two countries indicated the possibility of Japanese individual patients and doctor contents were useful as the cancer patients in the practical meaning. However, the current lists by the URL searching engines are including various noises such as SPAMs. To clarify the current status of the cancer information Webs in Japan and give the description for interventional frame work suitable information acquirement for the cancer patients and families with not only reducing the noises (ex. SPAMs and leading the commercial sites,) and too difficult and also enhancing the individual useful information from the searching results with words analysis and Web-mining variables.

♥ 正会員 東京工業大学大学院理工学研究科、情報通信研究機構、北陸先端科学技術大学院大学snakagaw@nict.go.jp
♦ 学生会員 北陸先端科学技術大学院大学情報科学研究科 s-kimura@jaist.ac.jp
♥ 非会員 情報通信研究機構新世代ネットワーク研究センター misumi@nict.go.jp
♦ 非会員 北陸先端科学技術大学院大学情報科学研究科 shimazu@jaist.ac.jp
♠ 非会員 東京工業大学大学院理工学研究科 yamaoka@ss.titech.ac.jp
♠ 非会員 東京工業大学大学院理工学研究科 ys@ss.titech.ac.jp

1. はじめに

がん患者にとってがんに関する情報は重要であり、生命予後にとって投薬や手術に匹敵する。最近Web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、がん患者やその家族にとってWebは重要な情報基盤のひとつとなりつつある。がん情報が他の医療情報に比べて盛んに流通するのは、治療法が確立し克服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５年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。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りたいという要求から少しでも多くの情報を得ようとする。医学分野でのWebデータ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SPAMの対象とな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ための必要条件について現状を調査しつつボトムアップに検討した結果を報告する。

2. わが国のがん情報流通の現状

2.1 各種がんの発生数と検索数の関係

わが国におけるがん情報提供の状態を図１に示す。これは平成11年におけるわが国の国際疾病分類（ICD-10[3],[4]）による部位別の死亡数（1年間の胃がん、肺がんなどの各項目での死亡数：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である。シンボル一つがんの種類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。年間の死亡数が1000件未満のものであってもヒット数は3000以上であった。しかし、Yahoo!等検索エンジンで「胃

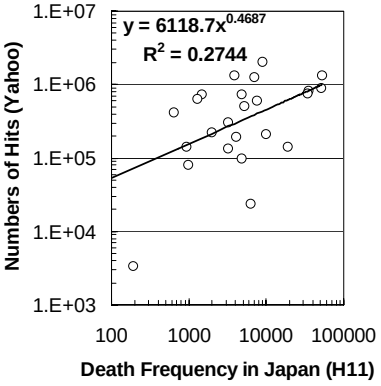

図１：H11 における各種がん死亡率とサーチエンジンによるヒット数の関係
Fig.1: Various Cancer Death (H11) and Searched Number of Hits

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。さらにGoogle, Yahoo!, infoseek等の汎用される検索エンジンは、現在一回の検索でのヒット数が数万、数十万あったことは示すが、実際にユーザが取得できるURLリストを1000以下に制限しており、これらのリストの母集団がいかなる質に基づくものかなどの検索をユーザレベルから行うことはできず、与えられたURLに一つ一つアクセスして情報取捨選択しなければならないのが現状である。これら現状をがん情報提供が早く始まった合衆国（US）とわが国の現状に関して予備調査し検討した。

2.2 Web 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)である。結果を表１に示す。Web データの収集には LinuxOS 上で 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が 50 に対して 201 であった。これら量の差は医療システムや予算の差が主な原因だが、がん患者にとって必要な情報が専門機関だけの情報発信では不足することがわかる。

2.3 コンテンツ提供者の特徴

2005 年 8 月、汎用される Web 検索エンジンを用い、NCC（日本）と NCI（合衆国）で共通に情報提供が行われている傷病名（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。結果を表２に示す。US では Cancer Net などの authorized institutions が Peer Review を行って発信しているコンテンツ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。US で Peer Review されたコンテンツが見かけ上多いのは、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）、国家戦略としてプログラムが存在することが主な理由である。個人による情報発信（例えば闘病記の類）は検索された URL リスト上位にはない。わが国では個人による、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などが特徴であり、これらコンテンツを有効に活用することによってコンテンツ量の不足と情報の多数性・多様性による妥当性の向上を狙える可能性がある。

表１：がん専門施設（日本 30，合衆国５）の Web ページ構成と各種 Web マイニングデータの比較（2005 年）

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

	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

表２：合衆国とわが国における胆管がんを検索キーワードとした場合の検索結果 100 の発信内容比較(2005.8 月)

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

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

3. コンテンツ内容のモデル化

3.1 がんに関するコンテンツ定型化の困難性

以上のことから、本研究では量の不足を補う必要があること、特に個々のがん患者と同じような状態にある患者自身の発信するコンテンツが専門機関からの情報を補完できる可能性に着目した。しかし、これら個人によって提供されるコンテンツは ne, co, or などからの発信であり、専門機関が ac, go, or 等の限られたドメインからの発信であるのに対して、発信元ドメインによって分類することは不可能である。このような場合、形態素解析により各単語の出現頻度からベイズ法、SVM 等の学習型分類器が用いられるが、教師データによる前学習が必要である。また医学的常識との整合性を確保できる保証はない。コンテンツ内容を数値化するために形態素解析を行うための専門用語辞書の作成も必要である。理想的には医学的オントロジーに基づいて決定木的に分類することであるが[5]、がんは疾患に関する診断・治療の標準化が行われておらず、定式化することが困難である。また情報提供者が個人であるため、例えば発信していた個人が死亡した時コンテンツへの到達性が失われることも問題である。また、がんの治療法は日々改変が行われており、作成されたコンテンツが単に古くなっても信憑性が失われることがあり得る。ただし、国家試験資格を持つ医療従事者は医師法によって情報を公表することに関する責任を問われるため、個人的な発信内容であっても情報の信憑性は高い。

そこで本研究では以下によって医学的整合性を保持した説明変数について検討することとした。(1)評価指標（CII: Cancer Information Index）を作成し、医学的見地から現状のサーチエンジンで得られる URL リストを分類・検討する。(2)各がんに関して一般に認められている Web ページ(今回、国立がんセンターの NCC-CIS を用いる）から専門用語辞書を作成し、単語空間から見たそれぞれの URL に対する語彙分析を行うことによって、異なり語数ならびに各 URL あたりの出現頻度を測定・検討し、異なり語集合を作成する。 (3)各種 Web マイニングデータの諸変数から CII を説明可能な変数を検索することとした。

3.2 コンテンツ評価指標の作成

Web サーチエンジンを用いて得られる URL リストの質的評価を大腸がん（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、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: 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。

以上得られた 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。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。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。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 を用いて非再帰的にダウンロードし対象とする HTML ファイルを固定した。それぞれの疾患別にダウンロードされたデータのコンテンツ量を Table 3 に示す。各 URL あたり約 100Kbytes、ファイル数 10 個であった。この大きさであれば、たとえばユーザからある URL に関して判定を求められたとしてもダウンロードに要する時間は長くなく（1Mbit/s で 0.8 秒）、問い合わせに対して逐次的に作業して結果を返答することが可能と考えた。

3.3 URL 分類の結果

表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。

図２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。傷病別に大腸がん、胃がんでは上位ほど商用サイトへの誘導や検索ノイズが高く、白血病では順位に伴って増加した。これより図１で予測したがんの発生数と検索ヒット数が相関する理由のひとつとして、頻度の多い胃がんや大腸がんに比べ、少ない白血病は SPAM のターゲット化していないことが考えられた。本結果からそれぞれの疾患名を検索語とする場合、平等に同数の URL リストを対象とすることは適切ではなく、これらノイズの存在を考慮してより多くの URL を対象として検索する必要性も示された。

表３：わが国における各がんの検索結果 100URL のコンテンツ量（2005 年 11 月）

Table 3: Web Data Contents Volume of 100 URLs of Each Cancer Keyword in Japan at 2005 Nov.

	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

表４：各がん検索結果（100 位まで）の分類結果
Table4: Result of URL Classification for five Cancers at Top 100 URL searched result

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

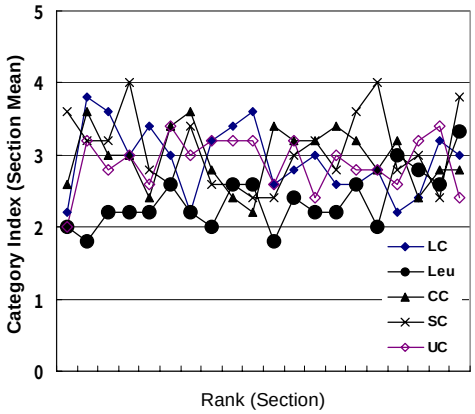

図２：各がんの検索結果ランクと分類スコアの５URL 毎の区間平均値

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

3.4 専門用語辞書作成と単語空間のモデル化

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

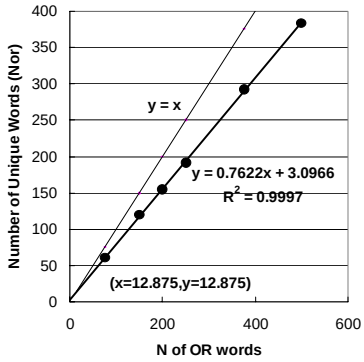

図３：5 種がんのトップ 100URL における出現頻度順の語数と異なり語数の関

Fig.3: Plot of 'Unique' vs. 'or' of Various Rank of 5 Cancers

し単語を切り出した。Leukemia、LC, SC, 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個の出現頻度を検討した。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（1URLあたり１回)。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図3に示す。本結果から、それぞれのがんにおけるNCC-CISでの単語の出現率が一般のURLでほぼ等しいことが示された。このことから一般辞書に加え、本専門用語群を形態素解析に用いて数値化したデータを用いてベイズ法を適応すれば医学的整合性を保ちうると考えた。

3.5 Web コンテンツ計測結果と CII の関係

さらに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(図4)。これよりHTML比がCII=2の分別に有効である事が示された。

3.6 がん情報 URL リスト適正化への検討

以上の検討によって、がん情報を提供しているURLからCII=2の群の選択のための要件として、それぞれの疾患により検索対象とするURLの総数を考慮する必要性、NCC-CISから作成した言葉集合を用いることの妥当性と医学的整合性の保持が可能である、Web コンテンツそのものの計測項目（Webマイニングデータ)の３つから問い合わせのあったURLを解析する事が可能であることが示された。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ところ[6]､教師データに対して良好な結果(正答率80%以上）を得た。しかし、分類根拠となった語（例えば、先生、私）の出現頻度と本研究で作成した専門用語空間とは異なっており、CII=2 と分類された結果に関してさらにサンプルサイズを大きくとって整合性を検討する必要がある。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、特定専門用語の出現確率に基づいた分類が処理も軽く実用的である。今後の検討課題とした。

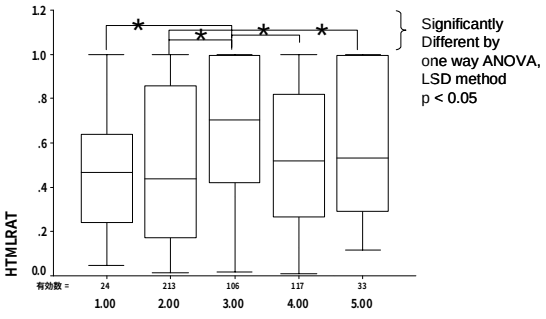

図４：CII と HTML 比の１元配置分散分析結果
Fig. 4: Result of One Way ANOVA for CII and HTML rate

4. まとめ

Webを介したがん患者のための情報提供URLリストの改善を目的として以下の結果を得た。

1. わが国におけるがん情報提供状態は専門機関ばかりでなく医師や患者個人による発信であることが特徴であり、がん情報資源の補完に有用である。
2. ５種のがんの検索結果について内容を分類した。それぞれの疾患でコンテンツの質の構成は異なっていた。
3. NCC-CISを元に作成した専門用語群によりURLベイズ分類に医学的整合性を付加できる可能性がある。
4. Webマイニングデータでは HTML 比が分類に関して有効であることが統計的に示された。

[謝辞]

本研究は情報通信研究機構運営費交付金,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。関係各位に謝意を表する。

[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 in Japanese 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 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

中川 晋一 Shin-ichi NAKAGAWA
情報通信研究機構主任研究員、京都大学大学院医学研究科終了、医師、博士（医学)、東京工業大学大学院博士後期課程在学中、日本データベース学会正会員

木村 俊也 Shunya KIMURA
北陸先端科学技術大学院大学博士前期課程在学中

三角 真 Makoto MISUMI
情報通信研究機構技術員、北陸先端科学技術大学院大学博士前期課程修了

島津 明 Akira SHIMAZU
北陸先端科学技術大学院大学情報科学研究科教授、工学博士

山岡 克式 Katsunori YAMAOKA
東京工業大学大学院理工学研究科助教授、工学博士

酒井 善則 Yoshinori SAKAI
東京工業大学大学院理工学研究科教授、工学博士