情報源としてのWWW

上田修一

1　はじめに

わずか10年も経たない間にWWW（World Wide Web）は，頻繁に利用されるメディアとして定着した。WWWは，本や雑誌などの印刷媒体，あるいはオンラインデータベースやCD-ROMなどの電子媒体とは異なった数々の特徴を持っている。作成者は，組織と個人が入り混じり，Webページは，環境と多少の技術さえあれば，誰でも作ることができ，知られさえすれば，不特定多数の人々に閲覧してもらえる。

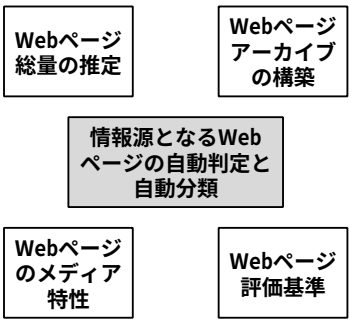
かつてWWWに対して，「落書きにすぎない」とか「空っぽの洞窟」であるといった批判があったが，見方によれば現在でも同様であろう。しかしながら，Web批判者が印刷媒体へのノスタルジックな想いを寄せている間に，日常生活の中での調べものにもっぱらWWWを使う人々が大量に出現している。インターネットの常時接続が一般的になれば，この傾向はますます強まるであろう。研究の上でもWWWが果たしている役割は大きい。特に専門書の翻訳といった調査を主体とする作業では，WWW上の情報源は欠かせない。単一のアクセス手順により辞書，百科事典，ハンドブック，統計，年表といったレファレンスブックに相当する資料を参照できるようになった。

WWWの中で，情報源として役立つ部分は，僅かであるに過ぎない。しかしWebページの急速な増加によって，情報源として利用できるWebページの絶対量も着実に増加している。

この研究は，情報源として役立つWebページを自動的に判断し自動分類を行なう方法を考案することを当面の目的としている。しかしながら，Webページの実態は，少しも明らかではない。インターネット利用が増加するとともに，Webページはどのようなメディアか，どのようにこれらを組織化するかに関する研究も増加しつつあることは確かであるが，まだ十分な蓄積がない。そのため，**図1**のような調査や研究も並行して行なっている。

Webページを研究対象とするためには，妥当な標本抽出が不可欠である。その基礎となるWebページの総量の推定は，米国NEC研究所のローレンスらによる調査[1]がよく知られているものの，推計のための手法には議論の余地が大きく残り，まだ確立してはいない[2]。また，この総量の推定をさらに進めて，特定時点におけるWebページの全体を保存する作業も必要である。こうしたWebアーカイブの構築は，米国では1996年からスミソニアン研究所や米国議会図書館，IBMによって「internet archive」[3]と呼ばれるアーカイブ計画が始まっており，1996年から1999年までで13.8テラバイトを収集，保存し，研究のために提供している。ただ，米国以外のWebページの収集については網羅性について疑

図１　Webページの検討課題



問が残る。

Webページのメディア特性の研究は，先にあげたようなWebページの作成，提供と利用の動機から行動にいたるまでのプロセスの解明を中心としたものである。それに，近年，商業的にも注目されているWebページの評価規準の設定という課題がある。

また，WWWにおける取り扱う単位として，ページ，ページ群，サイトのいずれが妥当かという議論もなされている，通常は物理的単位である１ファイルを１ページとして扱っているが，内容から見れば問題がある。ちなみに500ページを調査した結果，１ファイルで完結しているページは３割弱であるに過ぎなかった[4]。そのため，ページ群の自動判定を考える必要があり，いくつかの研究が行なわれている[5)6)7)]。

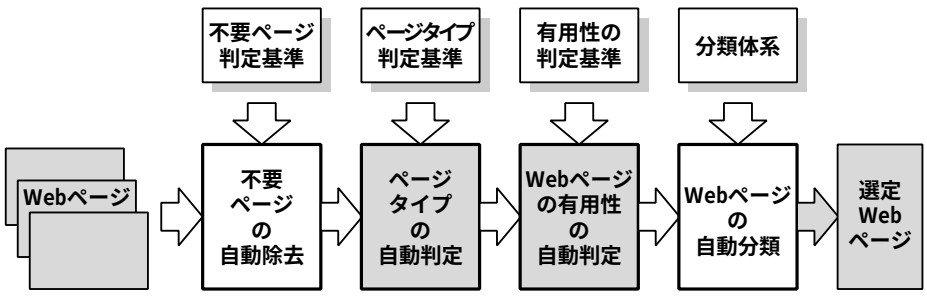
▶ 2　情報源となるWebページの自動判定

上記の図１の中央部分は現在行なっている作業である。これは，さらに**図２**のような個別の作業として示すことができ，収集したWebページ群から情報源となりうる分類されたWebページ群を取り出すまでの一連の手順を構成する。

これらは，

⑴　不要ページの除去
⑵　ページタイプの自動判定
⑶　情報源として有用なページの自動判定
⑷　自動分類

図２　自動判定・分類の処理手順



からなる。

Webページを網羅的に集めた場合，HTMLタグのみのページや未完成，あるいは死に絶えたページがかなりの割合で存在する。これらは，いくつかの判定規準によって自動的に除去できる。

自動分類はWebに向いた分類表の構築を含めて大きな課題である。

ここでは，⑴ページタイプの自動判定と，⑶情報源として有用なページの自動判定について述べる。

2.1 Webページの機能的，物理的特性

Webページの大きな特徴は，HTMLにより緩く構造化されたファイルであるという点であり，これには構造に関する情報を利用しつつ全文検索技術を用いた処理を，全てのページに適用できるという大きな利点がある。しかしながら実際には，HTMLにはW3C勧告の仕様書やDTD（Document Type Definition）等が存在し，タグ付けの規則は決まっているとはいえ，タグの使用法は作成者の好みで変えることができるような柔軟性を持っている。こうした柔軟性ゆえに，Webページの作成が広く普及したと考えることができるのであるが，次第に見栄えを重視したタグ利用が増えており，Webページの構造の分析が困難を伴うことは否めない。

また，Webページは不安定である。インターネットは分散型のネットワークであり，構造上，Webページの管理は各ページ単位で行なわれる。そのため，WWW上のさまざまな場所でページ単位あるいはサイト単位で発生と消滅が絶え間なく生じている。

この他，Webページの特色として常に言及されるのは，

⑴ リンク（ハイパーテキスト構造）

WWWは従来のメディアとは異なり，リンク機能によってハイパーテキスト構造となっている。

⑵ コミュニケーションの双方向性

フォームとCGI（Common Gateway Interface）などによって，WWWでは利用者が情報を受け取る一方ではなく，対話が可能になる。

⑶ マルチメディア

WWWではテキストだけではなく，静止画，動画，音声等のマルチメディアを扱うことができる。

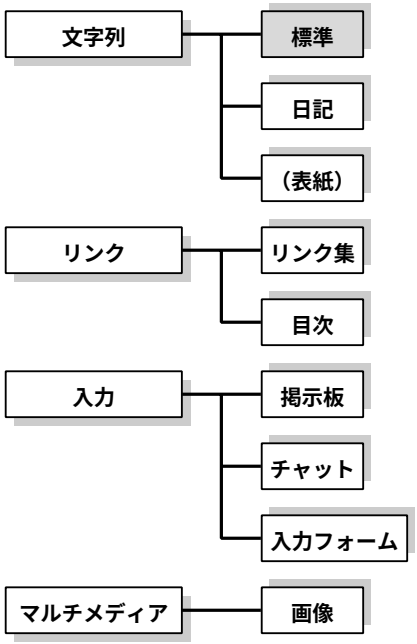
といった点である。

2.2 Webページタイプの自動判定[10)]

a ページタイプに関する研究

Webページのタイプをどう分けるかについては，既にいくつかの研究が行なわれている。例えば米国のハースらはページタイプとして⑴目次，索引（organizational），⑵参照，支援（documentation），⑶記事，論文(text)，⑷ホームページ（home page），⑸マルチメディア（multimedia），⑹入力フォーム（tool），⑺OPACなどの検索画面（database entry）の７種に分けている[8)]。一方，NECの福島らは，ビジネス用と個人用に大別したページタイプを設定し，これらの自動判定を行なっている。その成果はサーチエンジンNETPLAZAの「ページタイプサーチ」で使用されている[9)]。しかし，ハースらの分類は，わずか75ページをもとにしたものであり，Webの実態を反映しているとは言

図３　ページタイプ



い難い。福島らのタイプはもっぱら実用性を考慮したものである。

b　設定したページタイプ

ここでは，先にあげたWebページの特性をもとに**図３**のようなページタイプを設定した。

c　ページタイプの出現頻度

最初に，日本語のページを対象とし，ページタイプで分類し，出現頻度を調べた。

ロボット型サーチエンジンである「Info Navigator 疾風」は，ドメイン名による検索が可能である。そこでこのサーチエンジンを用い，ドメイン名毎にひらがな１文字を用いて検索し，全部で2000URLを抽出した。これらのページについて人手によりページタイプを判定した。最終的にページタイプを付与したのは，アクセスできなかったページや，１文字による検索によって検索されてしまう索引ページなどを除いた1,568ページである。この中で，上記のマルチメディアを除くページタイプ８種に該当する1,255ページを用いた。さらにこの中で1,000ページを訓練集合として特徴抽出に用い，残りの255ページを評価集合とした。

ページタイプの分布は**表１**のような結果となった。

d　ページタイプの判定規準と判定方法

ページタイプの判定規準に，Webページから得られる量的な指標やHTMLタグの出現頻度を用いることにした。

まず，ページタイプ毎に**表２**に示されるような量的指標とリンク数をカウントした。

次に，各ページに使用されているHTMLタグのページ毎の出現頻度を調査した。この

●表1 ページタイプの分布（1000ページ）

タイプ	ページ数	タイプ	ページ数
標　準	513	掲示板	99
目　次	108	リンク集	32
日　記	104	チャット	20
表　紙	104	入力フォーム	20

●表2 ページタイプ別の量的指標

	標準	目次	日記	表紙	掲示板	リンク集	チャット	入力フォーム	全体
文字数	14319.9	10512.1	14512.8	20180.9	49982.9	16894.0	19440.9	10196.5	18171.2
タグ数	579.4	666.6	572.1	1203.6	2142.1	1201.8	1175.5	787.2	843.7
コメント数	2.6	2.9	2.1	2.3	58.5	27.3	6.4	5.5	9.0
リンク数	17.1	99.0	27.3	25.3	101.3	178.3	45.1	104.9	43.7
HTMLリンク数	15.9	98.1	26.3	24.3	53.5	162.3	43.6	104.6	37.5
内部リンク数	13.2	86.0	16.3	21.1	43.1	94.2	36.6	102.6	30.0
外部リンク数	3.8	12.9	10.9	4.0	58.2	84.0	8.5	2.2	13.6
メールリンク数	0.4	0.4	0.3	0.5	47.3	15.5	0.4	0.2	5.5

Figure & Table

際に，ホームページ作成支援ソフトウェアによって自動付与されるHTMLタグがあるが，この影響をみるために，こうしたソフトウェアの使用状況を調べたところ，約26%のWebページで作成支援ソフトウェアが使用されていた。

各ページの量的指標とHTMLタグの出現頻度をもとに主成分分析を行ない，主成分とページタイプの関係を調べた。分析対象としたタグは，いずれかのページタイプにおいて50%以上の割合で使用されている16種のタグからbody, title, html, headといった基本構成タグを除いた計12タグである。

この結果，⑴リンク集と掲示板，目次は主成分分析で特徴がかなり表れること，⑵タグよりも量的指標のほうがページタイプの判定に利用しうること，などがわかった。

以上のような検討をもとにタイプ判定規準を考案した。

各タイプにおける特徴には大きなばらつきがあるため，例えば「外部リンク数＞50ならば，リンク集である」といった排他的な判定ルールでタイプ判定を行なうと，判定精度は極度に低下する。ここでは重み付けされた様々なタイプ判定ルールを用意し，その組み合わせから自動タイプ判定を以下のような手順で行なった。

① 対象ページの量的指標及びタグの出現頻度を解析する
② 各判定ルールを満たすかをチェックし，満たす場合，該当タイプに重みを与える
③ すべてのルールをチェックした後，最も高い重みが与えられているタイプを対象ページのページタイプ候補とする
④ 複数のタイプ候補があった場合には，訓練集合の出現頻度に応じた優先順位により一つのタイプに決める

その結果，126個のルールとその重みを設定した。以下にルールの例をあげておく。

タグ数/2 < リンク数 → 目次:+10
内部リンク数> 外部リンク数 → 目次:+10
内部リンク数<=外部リンク数 → リンク集:+10

これらは，”→”前の条件が満たされれば，後のタイプ別の重みを変化させることを表している。

●表３　自動判定結果

ページタイプ	再現率	精度
標準	78.4%	76.9%
日記	20.8%	12.2%
表紙	0.0%	0.0%
リンク集	25.0%	33.3%
目次	44.8%	68.4%
掲示板	61.1%	84.6%
チャット	60.0%	60.0%
入力フォーム	75.0%	60.0%

Figure & Table

e　評価結果

評価集合に対して自動判定手法を適用した結果は以下の**表３**の通りである。なお，評価集合中に「表紙」に該当するデータがなかったため再現率，精度ともにゼロである。

書式の定まっていないWebページにおいて量的指標のみからのタイプ判定は難しいが，ページタイプ判定後の処理の多くが標準タイプについて行なわれることから，今回は標準タイプの精度を上げることに重点をおいた。結果として標準タイプの識別に関しては，75%以上の再現率・精度を得ることができた。

2.3　情報源となるWebページの自動判定[11)]

a　Webページ評価の枠組み

既存のWebページの評価規準に含まれる評価項目と，Webページのアクセスを増やすために必要とされている項目を洗い出し，これらをまとめて，Webページの評価の枠組みを作り，そこから評価項目を導き，被験者にWebページを見せ，個々の評価項目の重要度をみることにした。

Webページを情報源として評価する規準と，アクセスを増やすための項目とを検討した結果，Webページの評価について作成者，利用者，物理的アクセス条項の三つの視点から整理した（**図４**）。

⑴　作成者の視点

当該Webページに対するアクセスを増加させるための評価項目である。これはWebページの構築に関わるもので，テーマの独自性，明確さなどが含まれる。

⑵　利用者の視点

利用者がWebページを閲覧する際に，何らかの情報を得るのに関わる項目である。内容の充実が大きく関与し，正確さ，速報性などが含まれる。

図４　評価の視点

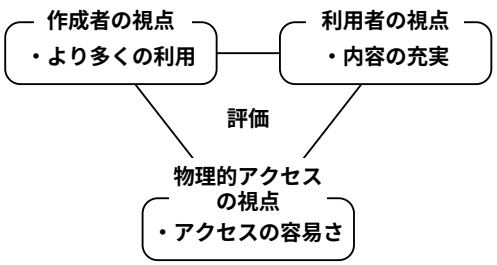


Figure & Table

●表４　評価項目

既存の項目	アクセスの増加	評価項目	視点
	量	量が豊富である。	作成者の視点
使いやすさ 一覧機能	操作性	見やすくするための工夫がある。	
「デザイン」	ページ構成 タイトル 統一性	ページのデザインがよい。 ページのタイトルが適切である。 ページ内のテーマが統一されている。	
リンク	リンク リンク集	他のページへのリンクが多い。	
対象利用者			
類似ものなし	独自性		
	コンセプト	テーマがわかりやすく明確である。 詳しい内容である。	利用者の視点
権威		作者に専門的な知識がある。 信頼できる作者である。	
正確さ		内容が正確である。	
新しさ		最新の内容である。 定期的に更新されている。	
文章の質		正しい日本語でかかれている。	
Smithの項目	アクセスの増加	評価項目	視点
アクセス性能	軽快さ		物理的アクセスの視点
	画像，音声等少		
	ページが小さい		
探索機能	データベース		その他
対話機能	掲示版等		
「評価の仕組み」，「支払う費用」			
	多言語，ソフトウェア		

Figure & Table

⑶　物理的アクセスの視点

当該Webページにアクセスするのコンピュータ利用環境への配慮などが含まれる。

評価項目を**表４**に示した。「既存の項目」は，各種のWebページの質の評価規準に含まれる項目，「アクセスの増加」はWebページ作成者に対してアクセスを増加させるために推奨されている項目である。

b　Webページの評価項目の調査

情報源とみなしうるWebページを選択し，これらはどのような評価項目と関連を持っているかを明らかにするために，被調査者に評価ページ群を閲覧させ，ページごとに「よい情報源」であるかどうか，さらに上記の14評価項目について５段階（5:強くそう思う, 4:そう思う, 3:どちらでもない, 2:そう思わない, 1:全く思わない）で判定させた。

調査の対象としたページ集合は，以下のような手順で収集した。

①Yahoo! Japanから約22万のURLを取得
②ロボットにより２レベル（と３レベルの１部）までの約500万のURLを取得
③無作為な5000URLを抽出し，画像等まで含めてダウンロード
④ページタイプ自動判定システムにより「標準」ページ群から1000URLを無作為抽出

●表５　「よい情報源」と評価項目間の相関

評価項目	相関係数
このページはよい情報源である。	1.0000
テーマがわかりやすく，明確である。	0.5798
信頼できる作者である。	0.4975
内容が正確である。	0.4963
詳しい内容である。	0.4878
見やすくするための工夫がある。	0.4594
ページ内のテーマが統一されている。	0.4308
量が豊富である。	0.4110
作者に専門的な知識がある。	0.3846
ページタイトルが適切である。	0.3730
ページのデザインがよい。	0.3520
正しい日本語でかかれている。	0.3290
最新の内容である。	0.3103
定期的に更新されている。	0.2829
他のページへのリンクが多い。	0.2338

●表６　語の出現状況

よい情報源（上位）			（下位）		
順位	頻度	語	順位	頻度	語
31位	229	経済	74位	83	笑
46	202	円	78	80	だっ
50	139	経営	82	78	けど
56	130	方式	84	75	今
68	113	料金	85	74	って
69	95	利用	86	72	でも
76	94	支店	88	71	ちゃん
79	90	サービス	90	70	コーラス
80	84	町	95	68	計
83	80	等	100	62	支部
85	79	施行	109	58	ノ
88	78	色	112	56	1K
91	76	条例	112	56	じゃ
93	73	接続	115	55	候
94	72	使用	116	54	いっ
100	70	必要	116	54	彼
111	63	基本	131	49	cards
116	61	大阪	131	49	return
119	60	高	131	49	来
121	59	郡	136	48	Case
123	58	管理	136	48	やっ
124	57	について	140	47	なかっ

Figure & Table

被調査者は，社会人，主婦，学生各３名の計９名であり，それぞれが500ページを判定した。計1000ページのうち500ページは６名，残りの500ページは３名で判定するように配分した。

c　「よい情報源」と評価項目間の関係

項目「このページはよい情報源である」と各評価項目との相関は**表５**の通りである。

さらに，前述の文字数，タグ数，リンク数，各タグの出現数などの量的指標の相関を分析したが，項目「このページはよい情報源である」との間には，直接的な相関は見られなかった。一方，「ページのデザインがよい」と画像数（0.26）「他ページへのリンクが多い」とリンク数（0.40），「量が豊富である」と文字数（0.25）などの間には，ある程度の相関が見られた（括弧内は相関係数）。

d　情報源となるWebページの判定方法

項目「このページはよい情報源である」と各評価項目との相関からみて，情報源となりうるページの判定に用いることができるのは，Webページの物理的，形態的な特徴ではなく，各ページのテキストであると判断される。そこでテキスト中の語の出現状況を調べた。

「このページはよい情報源である」の判定結果をもとに高得点のページから順位付けを行ない，上位の50ページと下位の50ページのテキストを対象として，「茶筅」を使って形態素解析を行なった。延べ単語数は，1,225,658語であった。よい情報源とされる上位のページ群と下位のページ群とにそれぞれ「単独に」出現する語のうち，記号や数字を除いたものが**表６**である。

Figure & Table

●表７　判定結果

評価対象	出現割合	
	上位語が多い	下位語が多い
良い（上位51位～100位）	78.0%	22.0%
良い（上位101位～150位）	80.0%	20.0%
悪い（下位51位～100位）	30.0%	70.0%
悪い（下位101位～150位）	44.0%	56.0%

e　判定結果

語の出現状況をもとに，被調査者による「このページはよい情報源である」の判定結果で高得点のページから順位付けで上位51位～150位，および下位から51位～150位のページのテキストを照合し，上位語と下位語の出現回数をカウントし，その比率を求めた。その結果を**表７**に示した。

こうした方法によりある程度，質的な側面の判定を行ないうることが判明した。

●参 考 文 献

1) Lawrence, S., Giles, C.L. "Accessibility of Information on the web". Nature. Vol.400, p.107-109（1999）
2) 安形輝. WWW調査におけるサンプル集合の収集法.　2000年度三田図書館・情報学会研究大会発表論文集．p.37-40（2000）
3) The Internet Archive: Building an 'Internet Library'< http://www.archive.org/>
4) 安形輝，上田修一他. WWWの実態調査とその方法, 1999年度三田図書館・情報学会研究大会発表論文集．p.17-20（1999）
5) 永藤拓宏; 遠山元道．ページ群への分割を利用したWWW検索エンジン．第9回データ工学ワークショップ（DEWS'98）,（1998.3.5-7）
6) 原田昌紀他. WWWページ間の階層構造の推定と検索システムへの応用. 情報処理学会研究報告（データベースシステム）Vol. 99, No. 39（99-DBS-118）, p.105-112（1999）
7) 石田栄美，上田修一他. 内容的なまとまりをもつWebページ群の自動判定. 1999年度三田図書館・情報学会研究大会発表論文集．p.21-24（1999）
8)Haas, S.W., Grams, E.S. Readers, Authors, and Page Structure　:A Discussion of Four Questions Arising from a Content Analysis of Web Pages. JASIS. Vol.51, No.2, p.181-192（2000）
9)松田勝志，福島俊一．　文書タイプ分類による問題解決向きWWW検索システムの開発と評価.　情報処理学会研究報告（情報学基礎）. Vol.99, No.20（99-FI-53）, p.9-22（1999）
10) 久野高志, 上田修一他. Webページのタイプ判定法. 2000年度日本図書館情報学会春季研究大会発表要綱．p.55-58（2000）
11) 上田修一他. Webページ評価の視点と基準. 2000年度三田図書館・情報学会研究大会発表論文集．p.33-36（2000）

（上田修一　慶應義塾大学文学部教授）