情報源となりうるWebページの判定

石田栄美（慶應義塾大学大学院）　　安形 輝(亜細亜大学)
久野 高志（作新学院大学女子短期大学部）　　上田修一（慶應義塾大学文学部）

情報源となりうる Web ページを自動的に判定する方法を考案するために，基礎的な調査を行った。まず， Web ページの既存の評価基準から三つの視点を明らかにし，15 項目の評価項目を得た。学生を対象とした予備実験によって，これら評価項目の妥当性を確認した後，社会人，主婦，学生の９名に対し，Web ページ 1,000 ページについて，各項目毎の評価と当該ページがよい情報源であるかどうかを判定させた。これらの作業によって，「情報源」となりうるページと強く結びつく諸要素が明らかになった。さらに，これらの要素を自動判定する方法を検討した。

Ⅰ　情報源としての Web ページ

Web ページは，初期の頃から情報源としての可能性が指摘されてきたが，Web ページの全体量が増えるにつれて，情報源と考え得るページ数が増加し，現在では，印刷体と並ぶ有力な情報源と考えられている。Web では，日常生活から研究活動にいたるまでの広い範囲の情報源を提供しているのでここでは「情報源」は，広い意味で用いる。Web ページには，全体に質が低いという批判がんされてきたが，量的な増大とともに質の高い Web ページの絶対量が増加したため，こうした批判はなされなくなっている。一方では，Web ページの評価に対して関心が集まりはじめており，電子商取引の分野では，サイトの評価を有料で行う試みも行われている。

情報源として有用な Web ページを自動的に判定することを目的して行っている一連の研究の中で，情報源となりうるページとこれに関連する評価項目の調査とその結果，および自動判定に利用しうる要素について提案を行う。

なお Web ページの評価を行う際に，評価対象としては，個々の Web ページ（ファイル）とし，サイト全体は考慮していない。また，ページタイプとその自動判定方法については，既発表 1)に基づき，テキストを中心とした標準ページのみを対象とした。標準ページはマルチメディアを除く Web ページのほぼ半数を占めている。さらに，日本国内の Web ページを主体とする。

II　Web ページ評価の枠組み

これまでの Web ページの評価基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とを洗い出し，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基準と，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。

(1)　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。

(2)　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。

(3)物理的アクセスの視点

当該 Web ページにアクセスする際の快適さに関わる評価項目である。ページのデザインに関連し，その量や利用者のコンピュータ利用環境への配慮などが含まれる。

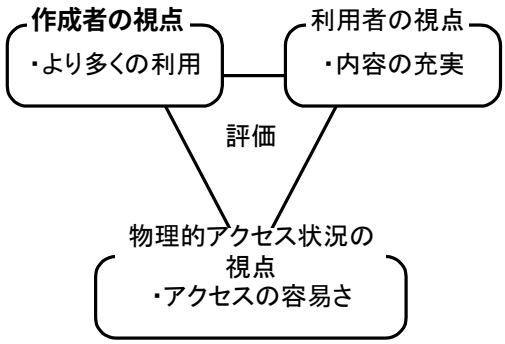


図1　　評価の視点

これら三つの視点のうち，作成者と利用者の視点に即して, 15 項目の評価項目を選定した。次に, 被験者に 30 ページの Web ページを見せ，これらの評価項目について評価を基準による評価実験を行った。被調査者は慶應義塾文学部図書館・情報学科の学生 26 名である。

評価基準として用いる項目とともに設問の一つに「このページはよい情報源である」を加えておき, この項目と評価項目間の相関を調べた(表1)。そして，被験者と対象ページを拡大し，本調査を行った。

表1　予備調査の結果

評価項目	相関係数
このページはよい情報源である。	1.00
作者に専門の知識がある。	0.52
詳しい内容である。	0.51
内容が正確である。	0.50
テーマがわかりやすい。	0.48
テーマが明確である。	0.48
信頼できる作者である。	0.47
正しい日本語で書かれている。	0.45
ページ内のテーマが統一されている。	0.35
ページタイトルが適切である。	0.32
量が豊富である。	0.26
ページのデザインがよい。	0.25
最新の内容である。	0.25
見やすくするための工夫がある。	0.25
定期的に更新されている。	0.17
他ページへのリンクが多い。	0.17

III　Web ページの評価項目に関する調査

1　調査方法

(1)　調査目的

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。

(2)　調査方法

a)　方法の概要

被験者に対象ページごとに「よい情報源」であるかどうかを判定させ，さらに各評価項目について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)で判定させる。

b)　評価項目

予備調査で用いた評価項目のうち，同じ回答パターンを示した「テーマがわかりやすい」と「テーマが明確である」をまとめ「テーマが明確でわかりやすい」とし，計 14 項目とした。

c)　対象ページ

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自動タイプ判定システムにより「標準とされた」ページ群から 1000URL を無作為抽出

最後の段階でページのタイプ判定を行っているが、これは「リンク集」、「掲示板」などを排除するためである。実際には、先行研究[1]において高い精度(76.9%)で「標準」ページを判定できたタイプ判定システムを使った。

d)　被調査者

被調査者は，社会人，主婦，学生各３名の計９名であり，その特性を表２に示す。

表2　被調査者の特性

区分	性別	年齢	インターネット利用歴
社会人１	男性	31歳	4年
社会人２	男性	32歳	5年
社会人３	男性	36歳	3年
主婦１	女性	34歳	3.4年
主婦２	女性	32歳	1.5年
主婦３	女性	35歳	4年
学生１	女性	18歳	1年
学生２	男性	20歳	2年
学生３	女性	19歳	0.3年

なお，各被調査者は 500 ページを判定したが，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ている。

２　調査結果

(1)　被調査者の属性による評価の違い

まず，被調査者の属性によって評価の違いの有無を調査した。図２は，同じ 500 ページを判定した６名の評価結果をクラスター分析した結果であるが，性別，年齢，インターネット利用歴による差はとりたててみられない。

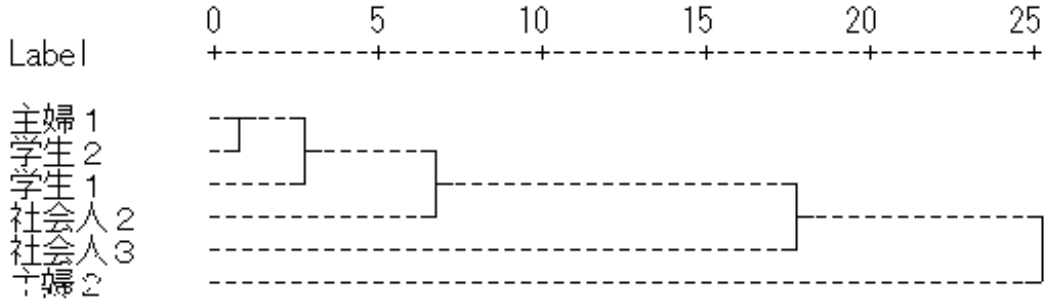


図２　　被調査者のクラスタ分析結果

(2)　被調査者間の「よい情報源」の判定の揺れ

項目「このページはよい情報源である」の一致の度合を見るために、各ページで判定された値の最大値と最小値の差を取った。値の差の平均は、調査者 3 人の 500 ページでは 1.29、6 人では 2.03 であった。

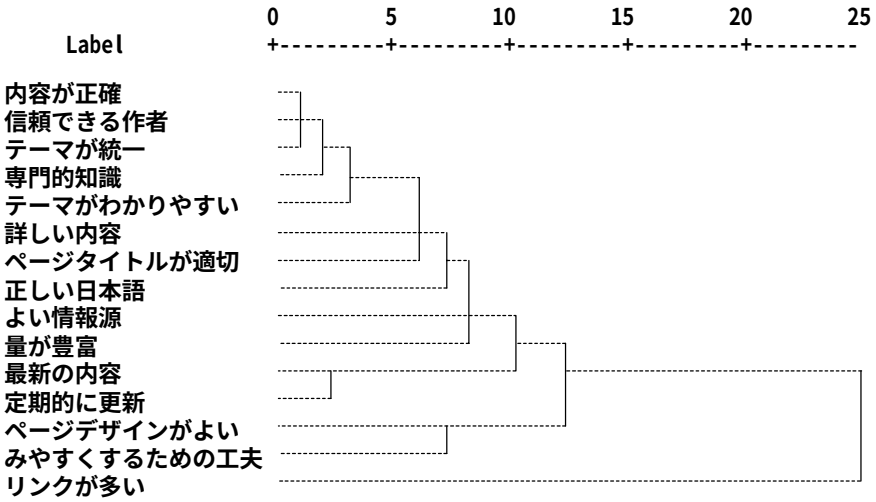
(3)　「よい情報源」と評価項目間の関係

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３のようになった。

また，各評価項目を最短距離法でクラスタ分析結果した結果を図３に示した。

表３　「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



(4) 各評価項目と定量的指標の相関

文字数、タグ数、リンク数、各タグの出現数などと定量的指標の相関を分析した。

項目「このページはよい情報源である」と定量的指標の間には、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）。

IV 情報源の判定方法

以上の結果を基に，よい情報源と判定されるWeb ページを自動判定するための手がかりを検討した。調査対象とした各 Web ページを「このページはよい情報源である」の判定結果をもとに順位付けし，上位のページと下位のページの特性の違いから以下のような手がかりを得た。

(1) ドメイン名

Web ページの著者は，評価に大きな影響を与えている。著者を自動的に判定するのは困難であるが，ドメイン名は一つの手がかりとなる。表４に示すように，go.jp を持つページは，上位（良い）の情報源となっている（表４）。

表４ ドメイン名

ドメイン名	情報源として 良い	悪い
ne.jp	30	32
co.jp	21	20
com	17	15
or.jp	14	19
go.jp	6	0
ac.jp	3	4
gr.jp	2	0
net	2	1
org	1	1
gov	0	1
その他	4	7
計	100	100

(2) テキスト中の語の出現条項

「このページはよい情報源である」の判定結果をもとに順位付けし，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った。述べ単語数は，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５である。

表５ 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
90	77	レ	112	56	じゃ
91	76	条例	115	55	侯
93	73	接続	116	54	いっ
94	72	使用	116	54	彼
100	70	必要	131	49	cards
109	64	A	131	49	return
111	63	基本	131	49	来
116	61	大阪	136	48	Case
119	60	高	136	48	やっ
121	59	郡	140	47	なかっ
123	58	管理	140	47	ら
124	57	について	143	46	なぁ

この他，「よい情報源」とされたページでは，見出しが使用されることが多いなどの特色が見られる。

以上の検討から「よい情報源」となりうるページの判断は，利用者の属性にはあまり依存せず，また，閲覧者の間ではほぼ一致すること，さらにその判断は，テーマや内容によって行われ，定量的要素との関係は薄いことが明らかになった。そして，テーマや内容に関わる一群の用語の出現状況が，「よい情報源」を自動判定するために利用できよう。

【引用文献】

1)久野高志; 安形輝; 石田栄美; 上田修一. 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)

2) 上田修一;久野高志; 安形輝;石田栄美. Web ページ評価の視点と基準. 三田図書館・情報学会研究大会発表論文集 2000 年度. p.33-36(2000)