Web ページの実態調査と自動判定
― ページタイプ判定と有用性判定を中心に ―

安形輝 (亜細亜大学)　野末道子 (鉄道総合研究所)　久野高志 (作新学院大学)
池内淳 (大東文化大学)　石田栄美 (国立情報学研究所)　上田修一 (慶應義塾大学)

【概要】 検索エンジンを使い収集した 3000 件の Web ページ集合を対象として、一般的なインターネット利用者である調査者 9 人に、ページタイプ・有用性・分類の項目に関して判定させた。結果としてはページタイプの一致度は高い、ページタイプ・カテゴリと有用性にはある程度の関連性がみられたなどが明らかになった。調査結果に基づいてページタイプと有用性の自動判定に関する簡単な実験を行っている。ページタイプ判定では、統計的な手法を導入することで自動的に判定ルールの組み合わせ方を決められる改良を行っている。有用性判定は先行研究である程度の精度を得られた手法に基づき、行っている。

1. 研究の背景
1.1 本研究における課題
本研究の目的は、Web ページの情報メディアとしての特性を明らかにし、その結果に基づいて Web ページの自動的な格付け、分類を行い、検索システムへ反映することにある。課題として、(1)総量の把握、(2)定量的特性の解明、(3)構造的特性の解明、(4)テキスト特性の解明、(5)ページタイプによる分類、(6)有用な Web ページの判定、(7)主題による分類、(8)寿命と永続性の把握などがあるが、ここでは、(2)定量的特性、(3)構造的特性、(4)テキスト特性、(5)ページタイプ、(6)有用性を取り上げる。

1.2 Web ページの実態調査の必要性
Web ページの検索エンジンや自動分類の領域で行われてきた研究では、アイディア主導で、Web ページの実態をきちんと調査せずに行われることが少なくない。一方で、Webmetrics あるいは Web 情報の有用性の調査[1]や質に関する調査[2]では、①調査対象数が少ない、②技術的な応用を想定していないなどの理由から、研究成果を Web の自動的な処理への応用することは難しい。有用性や質と関連し、Wathen らによる Web 情報の信頼性に関するレビュー[3]でも技術的な応用に関する研究は少ない。

そのような状況下で、本研究グループでは現在までに Web ページの自動的な処理を前提として、ページタイプ判定[4]、自動分類[5]、有用性判定[6]に関する調査と結果を応用し実験を行ってきた。しかし、これらはすべて個別のデータに対する調査であったため、ページタイプと有用性の相関など相互に関係付けた分析を行うことができなかった。そこで、今回は一つのデータに対して、三種類の調査を行っている。

1.3 自動判定の必要性
Web ページのタイプ判定に比べ、自動分類や有用性判定に関しては比較的多くの研究がなされてきた。このため、ここではタイプ判定の必要性について記述するに留めておく。

タイプ判定に関する研究が少ない理由としては、①タイプ判定は相対的に簡単な作業であり、研究面で追及すべきことがらが少ない、②タイプ判定を行うことができてもその応用範囲が狭い、ことなどが考えられる。

しかし、後述の調査結果からわかるように、分類判定や有用性判定に比べ、その結果のぶれが少ないため、機械的な判定を行いやすい部分である。また、タイプ判定は他の自動判定、例えば有用性判定とある程度の相関性を示すため、間接的に有用性判定などにも応用可能と考える。

2. Web ページの判定調査
2.1 調査環境
a. Web ページ集合

Webmetrics における対象ページの収集は Web 全体を代表させるという点では、包括的な Web ページアーカイブからの無作為抽出が考えられる。しかし、ここでは、インターネットの利用者がアクセスするであろう Web ページという点を考慮した上で、検索エンジンを使い収集した。具体的な手順は以下のとおりである。

① 5 つの検索エンジン(BIGLOBE, goo, FreshEye, infoseek, excite)を任意の意味をもたない文字と数字で検索し、1000 件以内の検索結果中をマージし、重複除去することで約 1 万件の URL を取得する
② 約 1 万件の URL から無作為に 3000 件の URL を抽出する
③ 3000 の URL についてページを再構成するのに必要な情報を画像等も含めダウンロードする

なお、調査対象ページの基本的な統計は表 1 のとおりである。検索エンジンを使った収集である

ため、以前の Webmetrics と比較し、若干文字数が増加している点以外はほぼ同じような特性をもつ集合となっている。

表 1　Web ページ集合の統計(3000 件)

URL	平均	標準偏差	最大値
バイト数	10000.1	28761.2	29960225
文字数	3468.8	14701.5	10392591
タグ数	222.2	422.6	665718
リンク数	18.4	33.9	55025

b. 判定者と判定ページの割り当て

Web ページの判定者は全部で 9 名であり、いずれも日常生活で Web ページを利用していることを条件に集め、その属性として、年齢は 20 歳代から 40 歳代であり、職種は会社員、主婦、大学院生などである。判定対象ページは3000ページであるが1000ページずつに分け調査者3名にページごとに判定をさせた。

c. 判定項目

判定者が Web ページに対して判定した項目はページタイプ、カテゴリ、有用性の3つから構成されている。

① ページタイプ

Web ページのタイプ分けには標準的なものが存在しない。そのため、タイプ判定を行うための枠組みとして、既存研究におけるいくつかのタイプ分けを検討した。

Haas らのページタイプ[7]は7種類［(1)目次、索引 (organizational) 、(2) 参照、支援 (documentation)、(3)記事、論文(text)、(4)ホームページ(home page)、(5)マルチメディア(multimedia)、(6)入力フォーム（tool）、(7)OPAC などの検索画面(database entry)］から構成されるが、少数のページを調査した結果に基づいて作られており、機能と内容を同じレベルで扱ってしまっている。また、松田らの研究[8]におけるページタイプは実用性を重視した結果、一貫性にかける。Koehler[9]による Web ページの永続性調査でもページタイプの観点を入れているが単に二種類に分けているだけである。

表 2　福島らのページタイプ

ビジネスユース	パーソナルユース
カタログ	
オンラインショップ	
FAQ	
リンク集	
調査報告	料理レシピ
求人案内	プレゼント
事例	教室・講座
イベント情報	アップデートプログラム

先行研究[4]では判定において曖昧性が少ないと考えられるタイプ区分として、図 1 に示される区分を設定した。これは機能面を重視し、できる限り体系化を試みることで、調査者がページを排他的に判定できることを目的としている。本調査にでもこのタイプ区分を用いている。

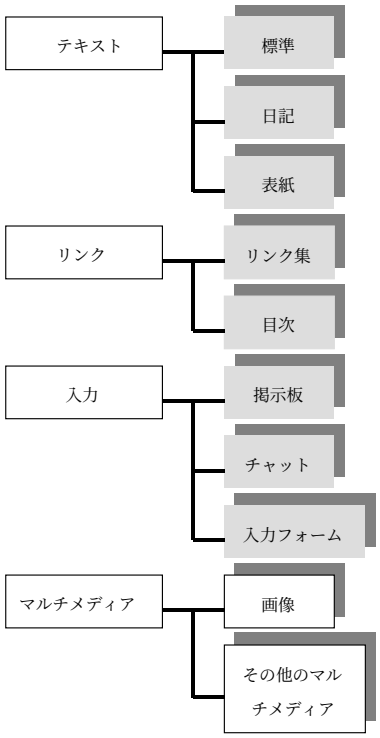

図 1　本研究でのページタイプ

② 分類カテゴリ

Web ページの主題は学術的なものから一般的なものまで多岐にわたるため、カテゴリは難しい。図書館における十進分類法を適用している先行研究もあるが、分類精度は低くなっている。ここでは、Web 環境においてもっとも利用されているディレクトリサービスである Yahoo!の第一レベルの13カテゴリとさらにその下の第二レベルカテゴリを用いた。

③ 有用性

有用性判定に関する先行研究[6]において、判定者間で有用性の概念を一致させることが困難であり、判定結果がばらついてしまったことが問題となった。そのため、ここではより具体性を持たせるため、「そのページを見て、有用がどうかを判断してください。あなたが Yahoo の査定者になったと仮定した場合に、そのページを登録するかどうかを判断して下さい。Yahoo は、質の高いページだけを集めているサイトです。」という形で、「採用する」「採用したいが問題あり」「どちらとも

表 3　判定の一致度

	ページタイプ		第一カテゴリ		第二カテゴリ(2,067)		有用性	
	件数	割合	件数	割合	件数	割合	件数	割合
三者一致	1236	41.2%	521	17.4%	358	17.3%	967	32.2%
二者一致	1470	49.0%	1876	62.5%	828	40.1%	1501	50.0%
三者不一致	294	9.8%	603	20.1%	881	42.6%	532	17.7%
合計	3000	100.0%	3000	100.0%	2067	100.0%	3000	100.0%

言えない」「残念だが採用しない」「採用しない」の五段階で判定させている。

2.2 調査結果

a. 判定者間の一致

調査環境に示したように一つの Web ページに対してそれぞれ 3 名が各項目について判定を行っている。ページタイプ判定では 3 人中 2 人以上が一致した割合（二者一致か三者一致）は 90%以上であり、一致度は高い。カテゴリ判定では第一レベルカテゴリでは 80%以上の一致度であるが、第二レベルまでの一致度は非常に低くなる。なお、第二レベルカテゴリでは第一レベルで分類不能とされたものを除いた 2067 件が総数となっている。有用性判定では 80%を若干切る程度の一致度になる。

b. ページタイプ

ページタイプの判定結果は表 4 のとおりである。集合中のページ数が 3000 であるのに、総数の合計が 9000 件であるのは判定者 9 人それぞれについて一致処理を行わず、合算しているためである。判定不能はページが不完全であるなどの理由でタイプ判定ができなかったものである。割合をみると、「標準」と判定されたページが半数近くを占めることがわかる。

表 5　判定されたページタイプ

	総数	割合
標準	3,879	43.1%
日記	239	2.7%
表紙	1,528	17.0%
リンク集	247	2.7%
目次	1,733	19.3%
掲示板	285	3.2%
チャット	80	0.9%
入力フォーム	119	1.3%
画像	468	5.2%
判定不能	422	4.7%
合計	9,000	100%

c. 分類

分類にはサーチエンジン Yahoo! Japan の分類を用いたが、ビジネスと経済(16.4%)、エンターテイメント(15.0%)、趣味とスポーツ(14.8%)、生活と文化（13.7%）が多数を占めている。なお、分類の判定結果については今後の研究で自動分類への応用を検討する。

d. 有用性

有用性の判定に関する結果は表 5 のとおりである。ここで真中の列「全ての判定値」は各判定者のすべての判定を単純に足したものであり、右の列「判定値の平均」は「採用する」を 5 から「採用しない」を 1 とした場合に各判定者の平均をとり四捨五入したものである。全ての判定を見た場合には「採用する」が最も多くなっているが、平均を取った場合「採用したいが問題あり」のレベルが最も多くなることがわかる。どちらの値をとるにしても、「採用する」「採用したいが問題あり」を有用性があるとまとめるならば、全体の 7 割程度の有用性があると判定されたことになる。

表 4　有用性の判定

	全ての判定値		判定値の平均	
	ページ数	割合	ページ数	割合
採用する	4,824	53.6%	816	27.2%
採用したいが問題あ	1,066	11.8%	1,298	43.3%
どちらとも言えない	906	10.1%	417	13.9%
残念だが採用しない	529	5.9%	405	13.5%
採用しない	1,610	17.9%	64	2.1%
判定不能	65	0.7%		
合計	9,000	100%	3,000	100%

e. 各項目のクロス集計

有用性とページタイプ・カテゴリ間の関係（全体の平均「採用する」57.9%）では、① ページタイプと有用性では、「目次」は「採用する」の割合が比較的高いこと（89.4%）、②カテゴリと有用性では、「政治」「自然科学と技術」は「採用する」の割合が比較的高いこと（84.2%、80.2%）、といった関連性が見られた。ページタイプとカテゴリ間の関係について関連性はあまり見られなかった。

3. Web ページの自動判定

3.1 実験環境

人手による判定済み Web ページ 3000 件からなる集合のうち、通し番号で No. 1～2000 までの

2000 件を学習用集合、No.2001～3000 の1000件を評価用に正解集合として使用する。各自動判定の評価は評価用集合を自動判定し、それが人手による判定と比較してどの程度の精度で行われているかによって行う。

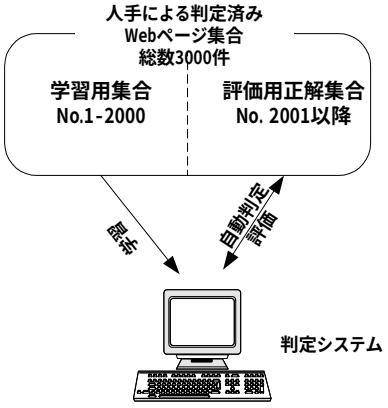

図 2 学習と評価手順

3.2 ページタイプの自動判定

先行研究[4]ではページタイプの自動判定について、データの量的指標から「文字数>5000」などのサブルールを洗い出しておき、学習用集合の統計的な特性を考慮しながら、それらのルールの重みを試行錯誤的に変更しつつ組合せる手法を用いた。この手法では、きめの細かい設定ができ、ある種のタイプに関しては非常に高い精度で判定することができた一方で、試行錯誤で組合せ方を決めるため再現性がなく、結果として出てきたルールセットと重み付けの合理的説明が難しいなどの問題があった。

ここでは前回の手法の改良としてルールの組合せに関して統計的な手法から自動的に重みを算出することを考案した。ページタイプの自動判定の具体的な手順は以下のとおりである。

① 学習用集合の量的指標から、ページタイプの特性を識別するであろうサブルール群を設定する
② 各サブルールについて、学習用集合中のページから、そのルールがあるページタイプにおいて成立する確率を求める
③ あるページタイプに関するサブルールの重みを②で求めた確率としてサブルール群の組合せ方を設定する
④ 評価用集合の各ページに対して、各サブルールを判定し重みを足していき、もっとも重みの高くなったページタイプをそのページタイプと判定する

3.3 有用性の自動判定

有用性の判定については、先行研究[6]で検討した複数の判定手法のうち比較的高精度が得られた手法を基本としている。今回の実験で具体的な判定は以下のような手順で行う。

① 学習用集合中において、「採用する」「採用するが問題あり」とされたものを「有用性がある」ページ、「採用しない」「残念だが採用しない」とされたものを「有用性がない」ページと設定する
② 各ページ群から形態素解析システムにより語を切り出し、語の出現頻度情報を得る
③ 「有用性がある」「有用性がない」ページ群両方に共通する高頻出語を除き、各ページ群に特徴的な高頻出語集合を作成する
④ 評価用集合中のページにおいて、特徴的な語集合中の語が出現するかによって有用かを判定する

4. 今後の展開

判定調査データを使い、自動分類を含め、さらに詳細な自動判定実験を行う。

【注・引用文献・参考文献】

1) 広田晃一. WWW の有用性について. *栄養学雑誌*, Vol.57, No.6, p367～371(1999)
2) Rieh, S.Y. “Judgment of Information Quality and Cognitive Authority in the Web”. *Journal of the American Society for Information Science and Technology*. Vol.53, No.2, p.145-161(2002)
3) Wathen, C.N.; Burkell, J. “Believe It or Not : Factors Influencing Credibility on the Web”. *Journal of the American Society for Information Science and Technology*. Vol.53, No.2, p.134-144(2002)
4) 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. 2000. p.55-58(2000)
5) 安形輝; 石田栄美; 久野高志; 野末道子; 上田修一. “WWW ページの自動分類 : NDC の分類体系と Yahoo のカテゴリを使った分類”. *情報処理学会研究報告* (99-FI-54). Vol.99, No.39, p.113-120(1999-05-17)
6) 石田栄美; 安形輝; 久野高志; 上田修一. “情報源となりうる Web ページの判定法”. 第48回日本図書館情報学会研究大会発表要綱. p.50-53(2000)
7) Haas, S.W.; Grams, E.S. “Readers, Authors, and Page Structure : A Discussion of Four Questions Arising from a Content Analysis of Web Pages”. *Journal of the American Society for Information Science*. Vol.51, No.2, p.181-192(2000)
8) 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)
9) Koehler, W. “Web Page Change and Persistence – A Four-Year Longitudinal Study”. *Journal of the American Society for Information Science and Technology*. Vol.53, No.2, p.162-171(2002)

10) Abbott, V.P. “Web page quality: can we measure it and what do we find? A report of exploratory findings”. Journal of Public Health Medicine, Vol.22, No.2, p.191-197(2000)