Web ページのタイプ判定法

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

増大する World Wide Web(以下，WWW)から，有用な情報源を自動的に選択する機能を持ったサーチエンジン開発のための基礎研究として，Web ページのタイプの分類と自動タイプ判定を試みた。Web ページのタイプについては「標準，日記，表紙，リンク集，目次，掲示板，チャット，入力フォーム」を設定した。量的指標やタグの出現頻度等の単純統計と主成分分析から導出したルールにより，ページタイプを判定する手法を考案し，判定実験から評価を行った。結果として，標準タイプに関しては 75%以上の判定精度を得ることができた。

1. Web ページの特色

インターネット利用が増加するとともに，インターネットにおける WWW とはどのようなメディアか，どのようにその情報源を組織化するか，に関する研究も増加しているが十分な研究が行われているとは言いがたい。

WWW 情報源を探すための方策を検討する前提として，WWW を他のメディアと比較し，どのような点に特徴があるか，物理的な側面，内容的な側面から整理した。

1.1 物理的特徴

a. 膨大な量

米国 NEC 研究所の Lawrence らは，Web ページの定量的な調査を行っているが，その結果によれば 1999 年 2 月現在では，世界中で約 8 億ページが存在している[1)]。

b. WWW の不安定性

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われる。結果として，WWW 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起こりうる。

c. 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。

d. 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。

e. 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。

1.2 内容的特徴

a. 書式が自由

HTML には W3C(http://www.w3c.org/)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。しかし，タグの使用法は各自の好みで変えることができるほどの柔軟性を持っており，結果として Web ページの作成が一般に普及したと考えられる。その一方で，見栄えを重視したタグ利用が増え，Web ページの機械的再編集や分析は困難を伴うようになった。

b. 主題は多様

WWW ではサーバ上に場所さえ確保すれば，誰でもが情報発信者になることができる。そのため，様々な主題を持つ情報が WWW 上に混在している。

c. 情報源として有用

WWW 上には一般的に有用なページは少ないと言われているが，その全体量が膨大であるために絶対的な量も多いと予測される。また，信頼性には欠けるが新しい情報も多く存在しているとされている。

d. ページの単位

WWW は，ハイパーテキスト構造をとること，1 ファイルの長さは決まっていないことから，1 ファイルを 1 ページとするページ概念は図書で言うところのページとは意味が異なる。

e. その他

チャットを中心に，コミュニケーションツールとしての利用が拡大している。これらの利用により得られた情報の中には，通常の Web ページからは得難い，最新かつ専門的なものが含まれる可能性がある。

The Judge method of Web page type.

Takashi, KUNO : Sakushin Gakuin University Women's College
Teru, AGATA : Asia University
Emi, ISHIDA : Keio University Graduate School
Shuichi, UEDA : Keio University

2. Web ページのタイプ判定

本研究では Web ページの自動タイプ判定を行うが，ここでは将来的なサーチエンジンとの関係からその必要性を検討する。

2.1 Web ページの組織化

膨大な WWW 上の情報源を組織化する方策としては，以下のようなものが存在している。そのもっとも代表的なものがサーチエンジンであるといえる。

a. ロボット型サーチエンジン

現在，ロボットを用いて Web ページを収集し，データベースを作成する方法が最も一般的である。

b. ディレクトリ型サーチエンジン

ディレクトリ型サーチエンジンは人手により Web ページを収集，評価し，探索手段を提供している。

c. リンク集

リンク集は，主題を限定して，人手により Web ページを組織化していると考えられる。

d. メタデータ

現在，ロボット型サーチエンジンは主として一次情報を検索対象としているが，検索精度向上のための二次情報としてメタデータの利用が考えられる。Web ページにおいては，HTML ヘッダ中のメタタグ（<meta>）にキーワード等を埋め込むことができる。正確さや統一性に問題があるが，当該ページの主題表現や探索のための手がかりをもっとも適切に与えることができる一人は作成者であることから，このタグ内容のメタデータへの反映が考えられる。

e. 目録作成

Web ページの書誌記述とアクセスポイントを人手により付与することにより，検索手段を提供しようとするものである。

2.2 Web ページ組織化の新しい形

現在は上記のような形で Web ページの組織化が行われている。しかしながら，人手による方法は，急速に増える Web ページに対処することがほぼ不可能と考えられる。また，メタデータのような作成者に依存する方法にはその強制力と正確さに疑問が残る。一方，ロボット型サーチエンジンは WWW の急速な拡大により，網羅性や検索効率の点で技術的な限界が見え始めている。

そこで，収集時に Web ページに対して様々な判定を自動的に行うことでこれらの問題を解決することを提案する。必要だと考えられる自動的な判定処理とは以下のようなものである。

a. ページ群の自動判定

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。500 ページを調査した結果では，1 ファイルで完結しているページは 3 割弱であった [2]。そのため，ページ群の自動判定を考える必要がある [3]。

b. 自動分類

ディレクトリ型サーチエンジンの利用度が高いことをみれば，分類からの探索にはかなりのニーズがあり，これに応える必要がある [4]。

c. 有用性の自動判定

内容の乏しい Web ページが多数存在しており，Web ページに対して格付けを含む質的評価を自動的に行う必要がある。

現在までの一連の研究から [2][3][4]，これらの処理を行うためにはその前提として Web ページを形式と内容からいくつかのタイプに分けることが必要であることが明らかになってきた。

2.3 Web ページのタイプ設定

a. ページタイプに関する先行研究

Web ページのタイプについては，既にいくつかの既往研究が行われている。

Haas らはページタイプとして(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)の 7 種に分けている [5]。

一方，NEC の福島らは，表 1 のように分け，これらの自動判定を行っており，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 [6]。

表 1　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

Haas 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ない。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1のようなページタイプを設定した。

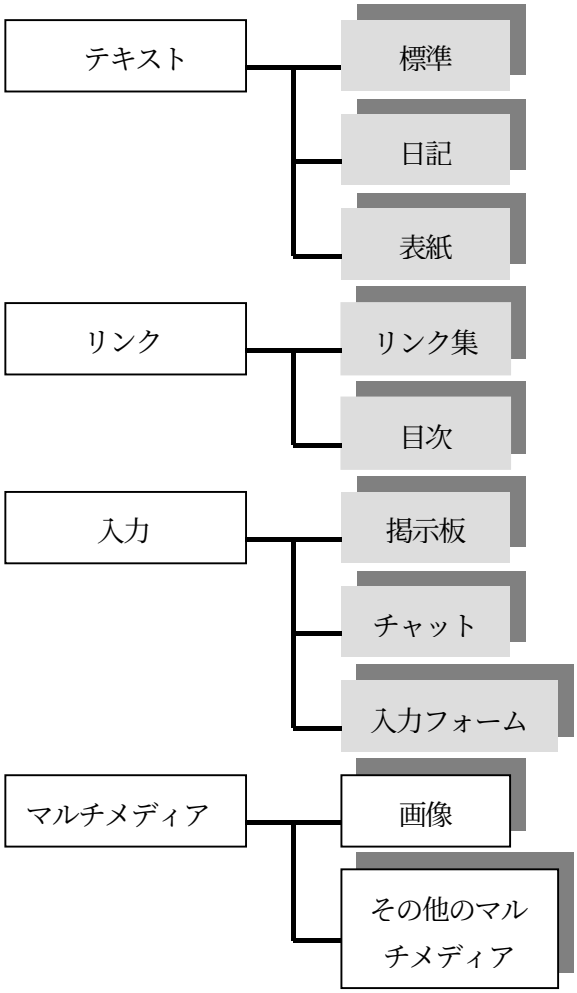

図1　ページタイプ

3. ページタイプの自動判定手法

Web ページを上記で設定したページタイプに分類する方法を検討した。量的指標や HTML のタグの使用頻度を用いる手法を考案した。

3.1 実態調査

a. 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てページタイプを確認し，出現頻度を調べた。

ロボット型サーチエンジン「Info Navigator 疾風」は，ドメイン名による検索が可能である。このサーチエンジンを用い，ドメイン名毎にひらがな1文字で検索し全部で 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る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，255 ページを評価集合とした。

b. 量的指標とリンク数

次に，ページタイプ毎に表3に示されるような量的指標とリンク数をカウントした。

表2　訓練集合のページタイプ

タイプ	ページ数
標準	513
目次	108
日記	104
表紙	104
掲示板	99
リンク集	32
チャット	20
入力フォーム	20
総計	1000

表3　量的指標平均値

指標	平均
文字数	18171.2
タグ数	843.7
コメント数	9.0
リンク数	43.7
HTMLリンク数	37.5
内部リンク数	30.0
外部リンク数	13.6
メールリンク数	5.5

c. HTML タグの出現頻度

各標本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。

表4は出現頻度が比較的高いタグとその頻度である。これらの HTML タグの出現とページタイプの関係を調査した。

表4　タグの出現頻度

タグ	ページ数	タグ	ページ数
body	952	hr	672
title	932	img	630
html	931	center	579
head	908	table	564
a	860	td	559
br	834	b	549
p	746	tr	549
font	744	meta	478

d. ホームページ作成支援ソフトウェア

ホームページ作成支援ソフトウェアによって HTML タグは自動付与される。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 ページで使用されていた。

e. 主成分分析

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く計 12 タグである。主成分分析においては相関行列を用いた。

この結果，(1) 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，(2)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手法を考案した。

3.2 タイプ判定手法

a. 判定手順

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うと判定精度が非常に悪くなってしまう。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う。

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

b. ルール作成

ルール作成には 1000 ページより構成される訓練集合を用いた。まず単純統計を行い，量的指標，タグ出現頻度から，人手によりタイプごとに大きく異なるものをルールとして抽出した。次に主成分分析を行い，結果として示された各主成分のうち，人手によるタイプと近いものの各指標をルールとして抽出した。

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，後ろのタイプ別重みを変化させることを表している。

3.3 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表５のとおりである。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，精度ともにゼロとなっている。

表５　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%

4. 考察

書式の定まっていない Web ページにおいて量的指標などからのタイプ判定は難しい。しかし，タイプ判定後の処理の多くが標準タイプについて行われることから，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。

謝辞

本研究は，佐伯文香氏によってタイプ判定された Web ページサンプル集合を分析対象として使用した。ご協力を感謝します。

【引用文献】

1) Lawrence, S., Giles, C.L. “Accessibility of Information on the web”. Vol.400, p.107-109(1999)

2) 安形輝，野末道子，石田栄美，久野高志，上田修一.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．1999．p.17-20(1999-10-16)

3)石田栄美，久野高志，安形輝，野末道子，上田修一．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．1999．p.21-24 (1999-10-16)

4)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)

5)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)

6)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)