データベースシステム　118－15
情　報　学　基　礎　54－15
（１９９９．５．１７）

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

安形輝　　石田栄美
亜細亜大学　慶應義塾大学大学院

久野高志　　野末道子　　上田修一
作新学院女子短大　鉄道総合技術研究所　慶應義塾大学文学部

【要旨】インターネットの発展とともにWebページ数は急速に増加し，国内だけで2000万ページに達しているとみられる。人手による選択と分類によるディレクトリ型では，増え続ける量を管理できないことが予想され，ロボットが網羅的にWebページを自動収集するロボット型では，内容の乏しいページが大量に出力される。そこでWebページを巡回するロボットに各ページの自動判断機能を持たせる必要があり，これは，不要ページの除去，ページ群の判定，有用性の判定，それに分類という手順になる。Webページの標本を元にその特徴を調査して，有用性についての目安を示した。さらに，文字列から形態素解析により抽出した語に相対頻度により重み付けする手法とn-gramにより文字列を抽出しベクトル空間型モデルによる情報検索を応用した手法によって，webページの自動分類を行った。なお，分類体系としては，『日本十進分類法』，Yahoo! Japanで用いられているカテゴリを用いた。さらに，比較のために，外部の分類体系としてCSJインデックスの分類表を用いた。

Automatic Classification of World Wide Web Pages

AGATA Teru　Asia University　(agata@asia-u.ac.jp)
ISHIDA Emi　Graduate School of Library and Information Science, Keio University
KUNO Takashi　Sakushin Gakuin Women's Junior College
NOZUE Michiko　Railway Technical Research Institute
UEDA Shuichi　School of Library and Information Science Keio University

The amount of World Wide Web (WWW) pages has grown dramatically over the last few years with the growth of internet. It is estimated that there are currently over 18 million WWW pages in Japan. In order to satisfy the requirement for new search engines for WWW pages, it is necessary to develop automatic mechanisms for the deletion of less important pages, the identification of identical pages, judgement of usefulness of pages, and classification. In order to classify WWW pages in Japanese, experiments of classification using NDC. Yahoo! categories, and CSJ index as classification scheme were conducted. We present two classification algorithms based on relative frequencies of terms and information retrieval technique using vector-space model.

I はじめに

様々な調査機関がインターネットのホスト数，ドメイン数を調査しているが，その報告にはかなりのばらつきがある。しかし，依然として高い伸び率を示していることは確かである。

調査の中で最もよく知られているのはネットワークウィザード社(Network Wizards)が年に2回行っているインターネットホスト数調査である。この調査結果[1]によれば，1999年1月現在のインターネットに接続されたホストの数は，表1のように約4，323万ホストであり，1998年7月(約3,674万ホスト)に比べて約700万ホスト増加している。日本(jp)は，169万ホストで，国別トップドメイン数では第1位となっている。この調査は「名前のつけられているIPアドレス」数を調べているものであり，直接の利用者数を示してはいない。実際の利用者数はこの数倍と考えられている。

表1 ネットワークウィザード社ホスト数調査

(単位:ホスト)

	1998年1月	1999年1月	伸び率
世界	29,669,611	43,229,694	45.7%
日本	1,168,956	1,687,534	44.4%

Webページ数を推計した例としては，1998年4月に *Science* 誌に掲載された S. Lawrence と C. L. Giles[2]によるものがよく知られている。彼らは，1998年はいめの検索可能なページ数を約3億2,000万ページとしている。

このデータと上記のホスト数の比率を用いて，国内のWebページ数を推計すると表2のようになる。

表2 Webページ数の推計

(単位:ページ)

1998年1月	1999年1月
320,000,000	466,240,000
13,000,000	18,720,000

すなわち，現在，世界中におおよそ5億ページ，日本国内には，2,000万ページ近くのWebページが存在するとみられる。

急速に増加するWebページは，これまでの出版物などとは異なり，何のコントロールもない状態で一定の書式で作られ公開されるメディアである。無内容なページが多いといった批判が根強いが，情報源として有用なページもまた数多く存在している。

WWWの発展の大きな要因となったのはサーチエンジンである。しかし，人手によって選択し分類をするディレクトリ型では，増え続ける量を管理できないことは容易に予想され，ロボットが網羅的に自動収集するロボット型では，内容の乏しいページが大量に収集，出力されるという問題を持っている。

以下では，こうしたサーチエンジンの可能性を検討し，特に自動分類機能に焦点をあてて，その方法について提案し，評価を行うことにする。

II 次世代サーチエンジンの要件

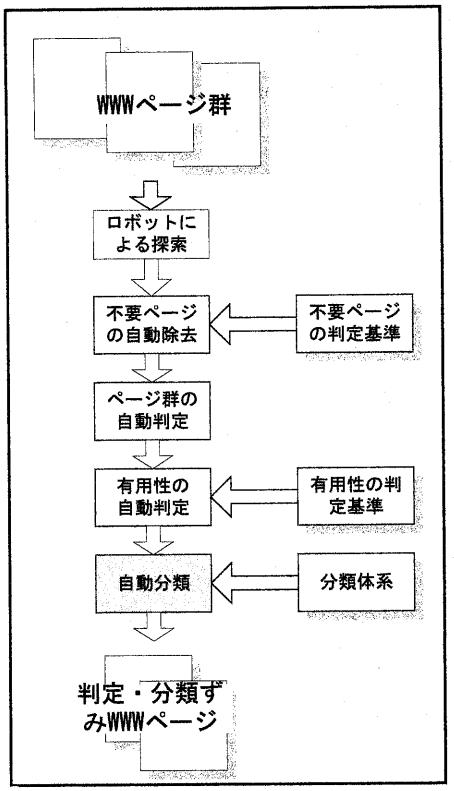
次世代のサーチエンジンでは，Webページを巡回するロボットに各ページを自動的に判断する機能を持たせる必要があろう。

こうしたサーチエンジンでは，図1のようなプロセスで，判定や分類を行うことになると考えられる。

1 不要ページの除去

不要ページの判定には下記のような基準を設定しうる。

- 画像のみ
- リンクが切れている
- "Not Found"など，Web特有のエラーメッセージである
- 自己紹介や日記などである
- チャットである(Q&A形式のものは除く)



2 Webページ群の自動判定

Webページには，1ページで完結するものと「ページ群」と呼ばれる内容的に関連のある一群のページの一部を構成するものがある。ロボットはページ群を自動的に判定し，そのトップページ（ホームページ）からページ群全体を取り出す必要がある。

ページ群の判定に関しては永藤拓宏と遠山元道が行った先行研究[3]が存在する。永藤らは意味的に最小のまとまりであるものと考えられるディレクトリ毎に，Webページを仮のページ群に分割し，その仮ページを，リンクの構造（張り方）と内容類似度比較を用いて6つのパターンに分け，分割・統合を再帰的に繰り返すことによって，ページ群を生成している。この研究における問題はディレクトリ構造を持っていないページ群を同定できない点である。

一つの手法として，ページ群の判定をトップページの識別で代表させることが考えられる。トップページとそれ以外のページを比較し，相互の差から以下のような要因をあげ，それらの要因に重みを与え，ある閾値を超えたものをトップページと同定する。

・ページ内のリンク先が同じディレクトリ，または，それ以下のディレクトリにある
・ファイル名がない，あるいは index.html（トップページでよく使われているファイル名）である
・リンク先ページ群の単語の出現頻度などから算出する類似度が高い
・ページのスタイル（レイアウト，ボディーカラーなど）が同一であるページ集合がリンク先に多くある
・連番のリスト中にリンクが数多くある

3. Webページの有用性の判定

次世代のサーチエンジンのロボットは収集時に，個々のページおよびページ群の有用性を判断する必要がある。

Ⅲ Webページの特徴

最初に，実際のWebページの内容に関する調査を行った[4]。

1. 調査方法

国内のページを入手するためドメイン名検索が行うことができ，jpドメイン登録数が多いAlta Vistaによりac.jp，co.jpなどのサブドメインから検索し，1000件のWebページの標本を抽出した。

2. 調査結果

a. Webページの形態的な属性

1）読込み可能性

集合中のWebページにアクセスし，読込むことができたページは935件であり、6.5%がアクセスできなかった。この結果はLawrenceらによる調査におけるサーチエンジン検索結果中の読込めないページの割合とほぼ一致している[2]。

2) 言語

自動判定により日本語とされたページは706 件(75.5%)であり，英語とされたページは 221 件(23.6%)であった。それ以外の 8 件は自動判定できなかったものである。

c. 文字数，画像数，リンク数

これらの集計結果を表3に示した。

表3 文字数，画像数，リンク数

	平均	最大	0個件数
文字数	943	114,869	2
画像数	4	140	289
リンク数	8	1,165	133
外部リンク数	2	1,165	690
上位リンク数	1	200	580
下位リンク数	5	291	241
メールリンク数	0	44	669

b. Web ページの内容的な属性

調査集合中の Web ページについては，さらに以下のような項目から構成される内容的な属性を複数の調査者により判断した。

表4 ページの作成者

	件数	割合
個人	310	36.5%
企業	336	39.6%
学校	70	8.2%
政府機関、非営利団体	89	10.5%
その他のグループ	44	5.2%

1) 作成者(表4)

作成者をそのページから判断し，個人，企業，学校，政府機関・非営利機関，その他のグループに分けた。「個人」と「企業」により作成されるページが 8 割を占める。

2) 内容のタイプ(表5)

内容のタイプとして，「教育的」，「学術的」，「一般的」，「娯楽的」に分類した。

3) 記載内容(表6)

記載内容を判断し，13のカテゴリに分類した。このカテゴリは経験的に作成したものである。出現する具体的な内容を示している。なお，一つのページが複数に分類されることもある。結果は表6のようである。

表5 内容のタイプ

	件数	割合
教育的	57	6.7%
学術的	88	10.4%
一般的	489	57.5%
娯楽的	216	25.4%

表6 記載内容

	件数	割合
論文	27	3.1%
企業の広告、紹介	225	25.5%
個人の創作	30	3.4%
自己紹介、日記	90	10.2%
物、事柄の説明(一般的)	142	16.1%
物、事柄の説明(学術的)	22	2.5%
チャット	25	2.8%
リンク集	154	17.5%
辞書	3	0.3%
メールのアーカイブ	18	2.0%
目次(本の目次などを含む)	15	1.7%
データベース	17	1.9%
その他	113	12.8%

4) 有用性の乏しいページと完結性

明らかに有用でないと判断されたページは約1割存在した。

また，ページの完結性を調査したが，およそ3分の1の 292 件(34.4%)ページが完結しておらず，ページ群の一部と判断された。

3. 国内 Web ページの特徴

調査の結果をまとめると以下のようになる。

1) 形態的な属性

・読み込うことのできないページの割合は1割弱である
・英語で書かれたページは22%である
・画像数の平均は3．9であり，ページ内に複数の画像が組み込まれている

2) 内容的な属性

・作成者は「企業」と「個人」で全体の約8割を占める
・「一般的」と判断されるものが過半数である
・収録内容は，「企業の広告，紹介」が多い
・完結性は，独立したページでないと判断されたものが全体の3分の1である

IV　Web ページの自動分類

Web ページの自動分類については，落合亮による先行研究がある[5]。これは，Web ページのテキストに対し，形態素解析を行い，形態素をもとに特徴ベクトルを与える方法で，Yahoo! のカテゴリに対し自動分類実験を行ったものである。また，徳田克巳らは，自動分類に「分類パターン」を用いる方法を提案しているが，「分類パターン」の一つとしてやはり Yahoo! のカテゴリを用いている[6]。

以下では，Web ページを対象として自動分類を試み，その手法と，結果を検討する。

1．分類体系

a. 日本十進分類法(NDC)

日本図書館協会『日本十進分類法』(NDC)は，図書等の分類に用いられており，分類時には，3桁以上の分類が与えられる。1997年の Japan MARC では表7のような分布になっている。

表7　NDC の区分と付与状況

NDC	点数	割合	3桁		全桁	
			種類	平均点数	種類	平均点数
0 総記	2,271	5.9%	40	56.8	230	9.9
1 哲学	2,644	6.9%	79	33.5	447	5.9
2 歴史	4,873	12.7%	62	78.6	541	9.0
3 社会	11,078	28.9%	84	131.9	1,714	6.5
4 自然	4,450	11.6%	89	50.0	870	5.1
5 技術	4,921	12.9%	98	50.2	905	5.4
6 産業	2,811	7.3%	75	37.5	731	3.8
7 芸術	4,327	11.3%	86	50.3	578	7.5
8 言語	920	2.4%	61	15.1	203	4.5
計	38,295	100.0%	674	56.8	6,219	6.2

b. Yahoo!のカテゴリ

Yahoo Japan では独自の分類体系を用い、Web ページの分類を行っている。実際に Yahoo Japan のサイトにアクセスし、ダウンロードできたデータを集計すると、各カテゴリに含まれるページ数と割合の分布は表8のようになる。NDC と比較するとカテゴリ間の格差が大きいことがわかる。

表8　Yahoo!カテゴリと付与状況

第一カテゴリ	ページ数	割合	第2カテゴリ
芸術と人文	8982	4.5%	30
ビジネスと経済	48389	24.0%	27
コンピュータ	7987	4.0%	17
教育	3111	1.5%	47
エンターテインメント	41403	20.5%	32
政治	1125	0.6%	11
健康と医学	4043	2.0%	37
メディアとニュース	4599	2.3%	26
趣味とスポーツ	6814	3.4%	18
各種資料と情報源	302	0.1%	19
地域情報	63566	31.5%	5
自然科学と技術	5681	2.8%	38
社会科学	1231	0.6%	26
生活と文化	4434	2.2%	50
計	201667	100.0%	383

c. CSJ のカテゴリ

Yahoo と同様に CSJ のカテゴリも独自の分類体系で Web ページの分類を行っている。CSJ のカテゴリないのリンクを分析すると表9のような分布になる。半分以上が趣味のカテゴリに入っていることが特徴的である。

表9　CSJ の分類と付与状況

大分類	ページ数	割合	中分類
ニュース	5351	3.0%	2
生活	19620	10.9%	2
各種情報	12014	6.7%	4
検索	14	0.0%	1
趣味	99734	55.7%	12
コンピュータ	12450	6.9%	3
ビジネス	21952	12.3%	3
学問	8061	4.5%	3
計	179196	100.0%	30

2. 自動分類に用いた手法

自動分類に用いた手法は2種類で表10にその概要示している。

a. 手法1：NDCを用いた自動分類

これは，図書データベースである Japan MARC に対する NDC の自動付与[7]で用いている手法である。ここでは，Web ページに対して NDC の上位3桁を各カテゴリと

し、NDC の上位 3 桁を Web ページに自動付与する手順を説明する。本手法は、すでに NDC が付与されているタイトルを用いて、各カテゴリに対する各単語の重みを求める学習フェーズと、対象文書をカテゴリに分類する分類フェーズからなる。

学習フェーズでは、まず、Japan MARC のレコードからタイトルと NDC のセットを抽出し、単語の各カテゴリごとの出現頻度から、単語に重みをつける。重みの計算

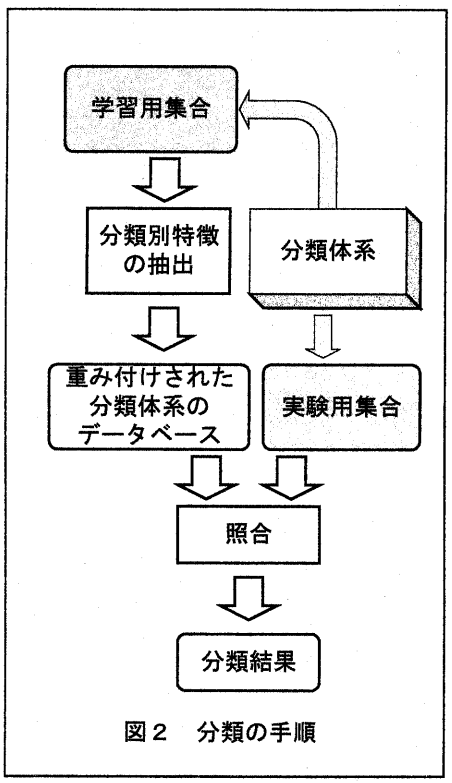


図２　分類の手順

は、カテゴリごとの単語の出現率を用いて行っている。

カテゴリ C_i における単語 t_j の相対出現率による重み w_{ij} は、以下の式で求める。

表 10　自動分類に用いた手法

	手法1	手法2
特徴の抽出方法	文字列から形態素解析により語を抽出	文字列からn-gramにより語を抽出
重み付けの方法	語の相対頻度	tf･idf法

$$w_{ij} = \frac{F_{ij}}{\sum_{i=1}^{N} F_{ij}} \log C/c+1$$

F_{ij} は単語 t_j のカテゴリ C_i における出現回数である。

分類フェーズでは、Web ページ中の単語 tj の各カテゴリ C_i における重みの積和を計算し、そのカテゴリ C_i が最大のカテゴリをその Web ページのカテゴリと推定する方法で行う。各カテゴリ C_i の特徴ベクトルを $c_i = \{ w_{i1}, w_{i2}, \cdots w_{ij} \}$ とし、Web ページの特徴ベクトルを $q_i = \{ w_{q1}, w_{q2}, \cdots w_{qj} \}$ と表わすと、Web ページの各カテゴリにおける重みは以下の式で求めることができる。

$$Sim(C_i, q) = \sum_{j=1}^{M} w_{ij} w_{qj}$$

Web ページとそれぞれのカテゴリに関して、$Sim(c_i. q)$ を求め、その値が最大であるカテゴリ C_i を Web ページのカテゴリと推定する。

b. 手法２：Yahoo を用いた分類

手法２による自動分類は、ベクトル空間型の情報検索手法を応用するものである。具体的には、以下の手順で行われる。

①学習集合中の文書を索引付けする
②該当文書を検索式として、それらの文書に対して検索を行う
③出てきた検索結果中の上位に出力された文書が含まれるカテゴリを付与する

手法２では、索引対象は、形態素解析による単語単位ではなく、記号を抜かした 2 文字の n-gram 手法で抽出した文字列としている。

抽出された文字列に対する重み付けは G. Salton による重み付け実験においてもっとも性能の高かった以下の式を使っている[8]。

学習文書中の文字列の重み付け：

$$w_{ij} = \frac{f_{ij} \cdot \log \frac{N}{n}}{\sqrt{\sum_{j=1}^{M} \left(f_{ij} \cdot \log \frac{N}{n} \right)^2}}$$

分類対象文書中の文字列の重み付け：

$$w_{ij} = \left(0.5 + 05 \frac{f_{ij}}{\max f_{ij}} \right) \cdot \log \frac{N}{n}$$

この式は、文書中での重要度を出現頻度 tf で、文書集合中での該当文字列の特定性を逆文献頻度 idf で表現し、それらの二つを組み入れた重み付けとなっている。情報検索における重み付け式で最も一般的なものである。

文書間の類似度の算出には重み付け式と同様に情報検索における一般的な類似度の算出式であるコサイン関数を使っている。

$$sim(d,q) = \frac{\sum_{j=1}^{M} (w_{dj} \cdot w_{qj})}{\sqrt{\sum_{j=1}^{M} w_{dj}} \cdot \sqrt{\sum_{j=1}^{M} w_{qj}}}$$

3. 実験対象データ

a. Web ページの収集

分類実験で、学習集合、実験集合として使う Web ページの収集は以下のように行った。

収集対象データとしては、CSJ あるいは Yahoo! Japan のカテゴリ構造中の HTML 文書と、それぞれのカテゴリ構造内の HTML 文書に URL が登録してあり、なおかつ、外部のホストに存在する HTML 文書としている。そのうち、a)エラー(File Not Found、Time Out)のページ、b)閉鎖・移転したページ、c)広告ページ（"ad. yahoo. co. jp"などの広告用サーバにある URL)、d)中身のないページとみなされるものは除去している。データ量が膨大なため外部ホストのリンクを読みこむレベルはトップページのみとしている。

収集したページからは不必要な HTML タグを除く処理を行っている。HTML 文書におけるタグは元来文書構造を表現するためのもので、理論上は自動分類への応用ができるはずである。しかし、現実のページでは HTML タグは文書構造というよりは文書レイアウトを整えるために使われており、元来の意味とは異なる使われ方をすることが多い。また、最近ではホームページ作成ソフトなどによって必要以上に大量の HTML タグが付与されているページも多い。従って、実際にはほとんどの HTML タグの自動分類への応用は難しく、不必要にインデックスの容量を大きくすることは望ましくないため、不必要な HTML タグをページから除去している。逆に除去せずに索引の対象とした HTML タグはメタタグ内で"keyword"や"description"を指定しているものである。

収集したページの所属するカテゴリは各カテゴリ間でその深度に大きくばらつきがあるため、各文書の所属カテゴリを第２カテゴリまでにデータのカテゴリ深度を揃えた。例えば、Yahoo Japan における「芸術と人文:パフォーミングアート:演劇:劇団」内の文書は上位の第２カテゴリである「芸術と人文:パフォーミングアート」に入ることになる。

b. 学習集合

手法１と手法２では学習集合が異なる。学習用として用いたデータ集合は以下の２種類である。

1) Japan MARC

手法１では学習集合は JapanMARC レコードの 1997 年の１年分（である 49433 件）から、900 番台を除いた 38242 件を用いた。

単語の抽出には、形態素解析 chasen を用いた。

2) Yahoo! Japan

Yahoo! から収集されたデータすべて201667件から、不適切なページと後述の実験集合1000件を除いた。処理を行った結果、約75000件となっている。

c. 実験対象集合

分類実験の対象となる実験集合は以下のような3種類となる。

1) Japan MARC

Japan MARC の 1998 年分から無作為抽出したデータ 1,000 件である。

2) Yahoo Japan

Yahoo Japan の収集されたページから不適切なページを除いたデータ集合から、無作為抽出した 1,000 件である。

3)CSJ インデックス

CSJ インデックスのカテゴリ「自然科学：宇宙学」、「自然科学：生物学」に含まれるすべてのページを対象として作成した。NDC を学習させた手法1と Yahoo Japan によって学習させた手法2の比較のために使われている。

4. 分類実験

分類実験は自動分類手法として手法1と手法2、学習集合として Japan MARC と Yahoo Japan、分類実験の対象集合として Japan MARC、Yahoo Japan、CSJ インデックスを使い、さまざまな観点で行われている。表3は今回行った分類実験についてまとめたものである。

【引用文献】

1) Network Wizards. Internet Domain Survey. [1999-04-10]<http://www.nw.com/zone/WWW/top.html>
2) Lawrence, S; Giles, C. L. "Searching the World Wide Web". Science. Vol.280, p.98-100(1988)
3)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. データ工学ワークショップ(DEWS'98), No.24, (1998.3.5-7)
4)石田栄美ほか. 情報源としてのWebページ. 1998年度三田図書館・情報学会研究大会論文集. 1998.
5) 落合亮. WWW ページの分類におけるテキストの特徴分析手法. 情報処理学会研究報告自然言語処理 118-14, p.85-90(1997)
6) 徳田克巳ほか. 分類パターンを用いた文書データの自動分類法. 情報処理学会研究報告自然言語処理 123-9, p.65-72(1998)
7) 石田栄美. [1999-04-17] http://www.slis.keio.ac.jp/~emi/classjpm.html
8) Salton, G.; Buckley, C. "Term-Weighting Approaches in Automatic Text Retrieval". Information Processing and Management. Vol.24, No.5, p.512-523(1988)

表 11　分類実験

手法	手法1	手法1	手法2	手法2
学習集合	3桁のNDCの付与されている1997年刊行の図書の書名。37,000件。	同左	Yahoo!の1999年4月現在の収録ページ。約120,000件。	同左
実験集合	3桁のNDCの付与されている1998年刊行の図書の書名1,000件。	CSJインデックスの学問とその下位カテゴリーに分類されたwebページ	Yahoo Japan!の収録ページから無作為抽出した1000件	CSJインデックスの学問とその下位カテゴリーに分類されたwebページ
結果	第1位で一致：52.9%，10位までで一致：77.4%	「宇宙学」：第一位のカテゴリに全体の33.3%が分類	第1位で一致：33.2%，10位までで一致：71.1%	「宇宙学」：第一位のカテゴリに全体の24.7%が分類