論 文

Webの情報検索
―ディレクトリサーチサイトの分類法―

清水由美子

Web上の情報検索に際し，初心者がよく利用するといわれる階層構造を持った検索サイトを対象に，その分類法を検討する．3つのディレクトリサーチサイトを，従来の情報分類法の一つである日本十進分類法を基準に比較した．最上階層の類似度，サブカテゴリ数とその分布，同一カテゴリ名の重複使用，分類の観点の交叉などを調査した結果，いずれのサイトも基本的に従来の階層構造を踏襲していたが，中には意図的に新しい分類法（交叉分類，カテゴリの複数箇所への位置付け等）を取り入れているサイトもあった．

キーワード：Webページの検索，ディレクトリサーチサイト，分類法

1 はじめに

ここ数年日本でもインターネットの使用者が激増している．その中でもWeb上の情報を有効に利用したいという要望は強い．しかしWeb上の情報はあまりにも膨大であり，自分の求める情報に的確に行き着くのはそう簡単ではない．また，得た情報がWeb上で得られる全てであったのかどうかという不安が常に付きまとう．情報利用に対する使用者の満足度は必ずしも高くないという調査結果もある（「Web上の情報に関する利用者の評価は厳しく，全体の80.5%が「情報を探すのに苦労している」と評価している[1]）．初心者にとっての情報検索の難しさについて，木村は学生を対象とした実験結果から，画面には「メディアの文法」があり，それが利用経験の少ない学生にとって障壁となっていると述べる[2]．

情報検索の際，初心者が多く利用するといわれるディレクトリサーチ（Webディレクトリを使った検索）をうまく利用するためには，「メディアの文法」を知ること，すなわち，それらディレクトリサーチサイトの構造を知ることも必要であろう．

本稿ではWebページの検索方法について，その種類と特徴を概観し，続いて3つのディレクトリサーチサイトの分類法を検討する．

2 Webページの検索方法

Webページの検索方法には大きく分けて次の三種類がある．

2．1 Webディレクトリ

Webページの情報をその内容に従ってカテゴリ分けし，カテゴリを辿ることで検索対象に行き着くように工夫された検索方法．基本的に階層構造（木構造・tree構造）の分類システムが採用されている．カテゴリはいずれも人が作ったものであり，掲載サイトの選択・分類も人手で行われている．各カテゴリが階層の中に位置付けられているため，そのカテゴリ名がどのような文脈の中で使われているかが容易に分かる．使用者の直感に従った検索が可能であるため，初心者には使いやすい検索方法である（1994年にサービスを開始したYahoo！[3]が代表格）．

2．2 検索エンジン

検索エンジン（ロボット）と呼ばれる検索用ソフトが，Web情報を蓄積しているサイトを自動的に順次訪問し，そのサイトの各Webページから自動収集したキーワード情報を自動的にデータベース化し，使用者はそのようにして自動生成されたデータベースを対象に検索を行う方法．キーワード検索と全文検索がある．キーワード検索は，検索対象となる情報のそれぞれにキーワードを付け，そのキーワードに対して検索を行うものである．ただし，キーワード付けはhtmlのタグから判断した特定部分について行われる．全文検索は，対象となるWebページの文章すべてに対して検索をかけるものである．ただし，大規模な文書群に対しては，検索実行時に全テキストを走査するのは現実的ではなく，テキストに出現するすべての単語に対して転置インデックスを作成し，これを利用して全テキスト走査と同等のことを実現する方法が取られる[4]．

検索エンジンによる検索では，ほとんどのサーチサイトで検索語に関連が高いと推定される順に結果が並べられる．時に何千，何万単位のサイトを拾い出してしまうため，そのうちのどれが本当に自分の欲しい情報であるのかの見極めが難しい．的確な情報を得るためには論理演算子（and,or,not）を利用して，情報の範囲を絞り込む必要があるが，キーワードを入力すること自体がバリアとなったり，演算子の適切な使用法が分からなかったりで，検索エンジンによる検索は初心者には敬遠されがちであるという[5]（日本のサイトではgoo[6]が代表格）．

SHIMIZU Yumiko
武蔵工業大学環境情報学部助教授

２．３　メタ検索エンジン

使用者の検索要求に対し，複数の検索エンジンを利用し，その結果（どの検索エンジンを使えばよいか）を返すシステム．自分自身がインデックスを持っているわけではなく，メタな知識を持った検索システムである．

３　ディレクトリサーチサイトの分類法

３．１　調査対象とそれぞれの特徴

日本語のディレクトリサーチサイトのうち Yahoo! JAPAN[7]，infoseek[8]，CSJ の厳選サイト集である CSJ What's Best![9]を対象とする．このほか，日本十進分類法[10]を分類法や分類カテゴリ検討の参考に用いる．

調査の対象とする階層と各カテゴリ数は次の通りである（なお，カテゴリは日々変化している．ここに挙げるカテゴリ数は 2000 年２月中の調査結果である）．

Yahoo! JAPAN：第１階層 14 カテゴリ，第２階層 431 カテゴリ，第３階層 4436 カテゴリ

infoseek:第１階層 18 カテゴリ,第２階層 271 カテゴリ,第３階層 1692 カテゴリ

CSJ What's Best!：第１階層 13 カテゴリ，第２階層 141 カテゴリ（第２階層から直接実際の Web ページにリンクされる）

日本十進分類法：第１次区分 10，第２次区分 100，第３次区分 941（1000 用意されている分類番号のうち，カテゴリ名が入らないものが 59 ある）

続いて，３つのディレクトリサーチサイトと日本十進分類法の特徴について概観する．

Yahoo！JAPAN：日本語での情報提供を目的とした世界中にいくつかある Yahoo!の一つ．第１階層はほぼ Yahoo！に対応するが，第２階層からは項目数などに違いが見られる．第２階層から直接実際の Web ページにリンクされているものもあれば（例：芸術と人文＞外国のアート＞実際のページ），６階層以上の深い階層構造を持つもの（例：自然科学と技術＞生物学＞植物学＞植物＞樹木＞林学，森林科学）もある．（「＞」で階層構造を表すこととする．）

infoseek:トップに 18 のチャンネル（第１階層）を設け，そのそれぞれがチャンネルリンク集（第２階層）を持つ．さらにその一つ一つのカテゴリがリンク集を持つ（第３階層）という構造．深さは第３階層までのものが主になっている．

CSJ What's Best!：厳選サイト集．第２階層から実際の Web ページにリンクされている．約 2000 件のサイトを掲載している．

日本十進分類法：学校図書館，公共図書館等で広く採用されている階層構造を持った分類法．第１次区分であらゆる文献・資料の扱う情報･知識を９区分し，それらに含まれないものと，二つ以上の区分を含むものを「総記」的な区分として設定する．第２次区分は，第１次区分のなかの一つの区分肢を，さらに９区分し，同じく９区分に含まれないものと，二つ以上の区分にわたるものを 10 番目の区分とする．以下，同様の考え方で，第３次区分，第４次区分，... 第ｎ次区分と分けていく[11]．

日本十進分類法は次のような分類原則を持っている[10]．

1)学術の体系に準拠

2)分類原則の遵守

(1)一貫性：一度に区別する原理は常に一つとし，無秩序な交叉分類を生じさせない，(2)相互排他性：区分された部分集合（種）はお互いに領域を侵さない排他的なものとする，(3)包括性：種の総和が被区分体（類）の全域を包含する，(4)段階性：区分は段階的でなければならず，飛躍を認めない，の４原則に従う．

3)分類項目名の明示

分類体系の各段階に列挙される分類項目名は，他の分類項目名と的確に区分される用語で明示されなければならない．単に用語が異なるだけでなく，その意味が明確で，紛らわしくないことも重要である．

4)過去への対応力を持つ

学者が提示する学術の分類は，現在の時点で最新の知識の状態だけに着目しているが，図書の分類では，過去の文化遺産である図書も分類する責務があり，現在では廃れてしまった文化の記録も収まる体系でなければならない．

5)将来の展望に備えて，新主題を収容できる余裕を持つ

図書の分類法は策定してから容易に変更できない固定性を持っており，それだけに将来の展望に備える余裕を確保すべきである．

３．２　最上階層の類似度

３つのディレクトリサーチサイト及び日本十進分類法の最上階層を対象に，四者の距離を見る．各サイト及び日本十進分類法の最上階層のカテゴリは次の通りである．

Yahoo! JAPAN：芸術と人文/ビジネスと経済/コンピュータとインターネット/教育/エンターテインメント/政治/健康と医学/メディアとニュース/趣味とスポーツ/各種資料と情報源/地域情報/自然科学と技術/社会科学/生活と文化

infoseek：ニュース&天気/マネー/ビジネス/キャリア/学び/コンピュータ/インターネット/カルチャー&ホビー/政治・法律・社会/クルマ/トラベル/スポーツ/エンタテインメント/ヘルス&ビューティー/住まい/グルメ&クッキング/くらしの情報/ショッピング

CSJ What's Best：ニュース/音楽・芸術/旅行・地域/ス

ポーツ/趣味/娯楽･グルメ/生活･医療/車/ショッピング/コンピュータ/教育・学術/金融・投資/ビジネス
日本十進分類法：総記/哲学/歴史/社会科学/自然科学/技術/産業/芸術/言語/文学

見た目の類似度を見るためにカテゴリ数，全文字数，１カテゴリ当たりの平均文字数，文字種（漢字・平仮名・片仮名・記号の延べ数と，全文字数に占める割合）を調べる（表１）．続いて，カテゴリ名の類似度を考察する（表２　もとになるサイトを縦に，対象とするサイトを横にとる）．カテゴリ名の類似は，他サイト等が全く同一のカテゴリ名を持っていた場合を１，一部が同じまたは類似のカテゴリ名を持っていた場合を 0.5 としてカウントする．なお，「エンターテインメント」と「エンタテインメント」，「クルマ」と「車」，「トラベル」と「旅」は同一カテゴリと扱う（例：Yahoo! JAPAN の「社会科学」に対し，日本十進分類法「社会科学」は１．infoseek の「ニュース&天気」に対し，CSJ What's Best の「ニュース」は 0.5）．例えば，Yahoo! JAPAN のカテゴリ名の自らのカテゴリ名との類似度は 14 であり，Infoseek のカテゴリ名との類似度は６である．ここで，各サイトのカテゴリ数が違うため，比較のために，自らのカテゴリ名との類似度を 10 とした場合の数値を（　）内に示した．これによると，Yahoo! JAPAN から見た日本十進分類法のカテゴリ名との類似度は 1.8 であるが，日本十進分類法から見た Yahoo! JAPAN のカテゴリ名との類似度は 2.5 である．

見た目には Yahoo! JAPAN と CSJ What's Best とが近いが，カテゴリ名では CSJ What's Best に対する infoseek が最も近い．

３．３　サブカテゴリ数とその分布

上位階層の各カテゴリに対して下位階層のいくつのカテゴリが対応しているか（サブカテゴリ数がいくつか）を見る．Yahoo! JAPAN の第１階層１カテゴリに対するサブカテゴリ数は平均 30.8（最少５カテゴリ，最多 50 カテゴリ）であり，第２階層１カテゴリに対するサブカテゴリ数は平均 10.3（最少０カテゴリ，最多 156 カテゴリ）である．以下，各サイトの第１～第３階層までのカテゴリ数，上位階層１カテゴリに対するサブカテゴリ数，サブカテゴリ数の分布（最少から最多まで）を表３にまとめる．なお，CSJ What's Best!は第２階層から実際のページへリンクされているため，第２階層１カテゴリに対するサブカテゴリ数等は求められない．

表３に見るように，Yahoo! JAPAN には第２階層から直接実際の Web ページに飛ぶカテゴリ(サブカテゴリ数 0)

表１　見た目の類似度（%は小数点第２位を四捨五入）

	カテゴリ数	全文字数	平均文字数	漢字（%）	平仮名 (%)	片仮名 (%)	記号（%）
Yahoo! JAPAN	14	86	6.1	41 (47.7)	9 (10.5)	36 (41.9)	0 (0)
Infoseek	18	94	5.2	10 (10.6)	7 (7.4)	71 (75.5)	6 (6.4)
CSJ What's Best	13	55	4.2	25 (45.5)	0 (0)	24 (43.6)	6 (10.9)
日本十進分類法	10	24	2.4	24 (100)	0 (0)	0 (0)	0 (0)

表２　カテゴリ名の類似度

	Yahoo! JAPAN	infoseek	CSJ What's Best	日本十進分類法
Yahoo! JAPAN	14(10)	6(4.3)	5(3.6)	2.5(1.8)
Infoseek	6(3.3)	18(10)	9(5)	0(0)
CSJ What's Best	5(3.8)	9(6.9)	13(10)	0(0)
日本十進分類法	2.5(2.5)	0(0)	0(0)	10(10)

表３　各階層のカテゴリ数，及び上位階層１カテゴリに対するサブカテゴリ数とその分布

	第１階層カテゴリ数	第１対第２サブカテゴリ数(平均)	サブカテゴリ数分布	第２階層カテゴリ数	第２対第３サブカテゴリ数(平均)	サブカテゴリ数分布	第３階層カテゴリ数
Yahoo! JAPAN	14	30.8	5～50	431	10.3	0～156	4436
Infoseek	18	15.1	5～25	271	6.2	0～47	1692
CSJ What's Best	13	10.8	5～20	141			
日本十進分類法	10	10	10	100	9.41	1～10	941

もあれば，第２階層１カテゴリ当たり 156 というサブカテゴリを持つものもあり，サブカテゴリ数のばらつきが大きい. 100 以上の大きなサブカテゴリ数は次のような分類法によるものである.「地域情報（第１階層）＞世界の国と地域（第２階層)」の下位階層には「アイスランド」「アイルランド」「アフガニスタン」といった国の名前が並び，これらの国名を合わせると 140 カテゴリとなる. infoseek にも「トラベル（第１階層）＞世界の地域情報（第２階層)」のような類似カテゴリがあるが，このサブカテゴリは 11 である. こちらの分類は「ヨーロッパ」「北欧」「中東」のように，かなり大きなレベルで行われている．また，Yahoo! JAPAN の「健康と医学（第１階層）＞病気，症状(第２階層)」は 129 のサブカテゴリを持つ. ここには「インフルエンザ」「肝炎」「咽頭癌」といった多くの病名が並べられると共に「肝臓病」「癌」のような，その上位レベルのカテゴリ名も同列に置かれている.

３．４　同一カテゴリ名の重複使用

一つのカテゴリ名が同一階層内で，あるいは別階層で複数回使用されることがある．CSJ What's Best!の第１階層カテゴリ名「コンピュータ」は，同じく第１階層「ショッピング」のサブカテゴリ名としても用いられる．また，infoseek でも「キャリア＞資格取得」「学び＞資格取得」のように同一カテゴリ名が別の上位階層のサブカテゴリ名として使われている．日本十進分類法の第１区分カテゴリ名も第２区分以下で用いられる（例　第１区分「１哲学」，第２区分「10 哲学」，第３区分「100 哲学」）. ただし，これらはいずれもカテゴリ名が同じであるからといって同一のサブカテゴリを持つわけではない. What's Best!の第１階層カテゴリ「コンピュータ」のサブカテゴリは「全般/マック/PCWindows/UNIX/ハード一般/ソフトウェア/コンピュータ雑誌/周辺機器」であり，「ショッピング」のサブカテゴリ「コンピュータ」は実際の Web ページにリンクされている.

カテゴリ名の重複使用の度合いをそれぞれのサイトで見た.（表４　例：Yahoo! JAPAN の第１階層カテゴリ名は，同一階層では重複使用されないが，第２階層では延べ 18 回出現する．Yahoo! JAPAN 第２階層カテゴリ名は，第１階層で延べ 18 回，同一階層内では延べ 561 回の重複使用がある.）

ここで特徴的なのが，Yahoo! JAPAN における同一カテゴリ名重複使用の多さである．これは Yahoo! JAPAN が次のような階層構造のルールを持っているためである[12].

> あるカテゴリやサイトを配置する際，一つのカテゴリに置くだけでは不十分な場合があります．「スポーツニュース」のあるべき場所は「趣味とスポーツ：スポーツ：メディアとニュース」でしょうか，それとも「メディアとニュース：スポーツ」でしょうか．どちらも可能性があり，どちらか片方だけでは不十分です．このような場合，Yahoo! JAPAN では「カテゴリリンク」という方法を用いて，一つのカテゴリを複数のカテゴリに表示できるようにしています.

例えば，「エンターテインメント＞ゲーム」のサブカテゴリと，「趣味とスポーツ＞ゲーム」のサブカテゴリは，共に「アーケード/囲碁/イベント/インターネットゲーム/…」である．これは，一つのカテゴリにアプローチするのに複数の方法がある，または一つのカテゴリを階層構造の複数箇所に置くということであり，他のサイト及び日本十進分類法には見られない分類法である．Yahoo! JAPAN では，積極的にこのような分類法を用いているため，同一カテゴリ名の重複使用が多くなっている.

３．５　分類の観点

Yahoo！JAPAN には，観点の異なる分類が同一ページに掲載されている場合がある．例えば「芸術と人文＞人文＞文学＞販売」のサブカテゴリには「イベント/書店/タイトル/取次/本の検索サービス」等が並ぶ．また同一ページにライン一本で分けられて，「SF，ファンタジー，ホラー/アダルト/エンターテインメント/音楽/技術/教育/ミステリ/料理」等，本のジャンル別カテゴリが並んでいる．このように，一つの事柄についての異なった観点からの分類を同時に掲載しているページがいくつか見られる.

４　おわりに

４．１　まとめ

今回調査した３つのディレクトリサーチサイトは，階層構造を持つという点で従来の分類法の一つ，日本十進分類法を踏襲している．infoseek，CSJ What's Best! の

表４　同一カテゴリ名（出現階層と延べ数）

	Yahoo! JAPAN		infoseek		CSJ What's Best!		日本十進分類法	
	1	2	1	2	1	2	1	2
1	0	18	0	5	0	2	0	8
2	18	561	5	8	2	10	8	0

第１階層はそのカテゴリ名からかなり斬新な印象を受けるものの，基本的に分類法においては日本十進分類法を大きく逸脱していない．

しかし，日本十進分類法のカテゴリ名に最も近い第１階層カテゴリ名を持つYahoo! JAPANは，「1度に区別する原理を常に一つとする」[10]といったその分類原則に反して交叉分類を採用したり，一つのカテゴリを複数箇所に位置付けるなど，カテゴリ分類法で意図的に新しい方法を取り入れている．

４．２　今後の課題

今回の調査では，それぞれのサイトがカバーするサイト数の規模が違うため，必ずしも正確な比較ができたとは言えない．今後，同程度の規模のディレクトリサーチサイトを対象とした調査を行い，より正確な対比を試みたい．

また，各サイトについて，上位階層と下位階層との意味関係の分類を行う等，更に詳しい分析によりそれぞれのサイトの分類法の特徴をいっそう明確にしていきたい．

Yahoo! JAPANでは，従来にない新しい分類法を取り入れていた．こうした新しい分類法が，本当に初心者にとって検索しやすいものであるかどうかの検証も今後の課題である．例えば，いずれのディレクトリサーチサイトも，自分が今居る位置を様々な仕方で表示している．Yahoo! JAPANではページ上部左に「ホーム＞健康と医学＞女性の健康」のような形で現在位置を示しているが，一つのカテゴリを複数箇所に位置付けた場合，次のような事が起こる．「ホーム＞エンターテインメント＞芸能人，タレント＞アーティスト」と辿ると，次の段階でのページの位置表示が「ホーム＞エンターテインメント＞音楽＞アーティスト」となる．自分が辿った通りの道筋が表示されないことから，使用者が戸惑いを感じる可能性はないか等，実際の検索行動の観察なども併せて行いたい．

参考文献

[1] 臼澤基紀・新垣紀子・野島久雄・石崎雅人：“WWW検索行動における「戻る」行動と検索方針の変化との関係，”情報処理学会ヒューマンインタフェース研究会発表資料，99-HI-83，pp.61-66，1999

[2] 木村忠正：“「次の10件」に気付かない学生たち－メディアの文法とネットワークへの受動的意識構造－，”日本語学九月臨時増刊号，第十七巻第11号，明治書院，pp.188-203，1998

[3] Yahoo！，http://www.yahoo.com/

[4] 長尾真・黒橋禎夫・佐藤理史・池原悟・中野洋：岩波講座言語の科学９　言語情報処理，岩波書店，1998

[5] 新垣紀子・野島久雄：“電子メディア社会における人の情報検索プロセス，”CMCC研究会第１回シンポジウム論文集，pp.3-12，1999

[6] goo，http://www.goo.ne.jp/

[7] Yahoo! JAPAN，http://www.yahoo.co.jp/

[8] infoseek，http://japan.infoseek.com/

[9] CSJ What's Best！，
http://www.csj.co.jp/whatsbest/

[11] 柴田正美：資料組織概説，日本図書館協会，1998

[10] もり・きよし原編　社団法人日本図書館協会：日本十進分類法新訂９版本表編，社団法人日本図書館協会，1995

[12] Yahoo! HOW-TO？，
http://www.yahoo.co.jp/docs/howto/chapters/8/6.htm