修 士 論 文

地域情報を対象としたウェブディレクトリの自動生成

指導教官　佐藤 理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

大槻 洋輔

2000 年 2 月 15 日

要 旨

本論文では, 地域情報を対象としたウェブディレクトリと, それを自動生成するシステムを提案する. 本システムの中心技術は, 情報源となる地域サイトの自動収集と, 地域情報ページの自動分類である. 地域サイトの自動収集では, 地域サイトに見られる典型的な URL パターンを利用して, 既存の地域サイトのリンク集を発見し, そこから地域サイトの URL を収集する. この方法により, 日本の全地域 (3427 自治体) の 80%以上に対して, 情報源となる地域サイトを収集することができる. 一方, 地域情報ページの自動分類では, 収集した地域サイト内のページを 8 種類のカテゴリに分類する. この分類は, それぞれのカテゴリに固有な表現が, ページのタイトルやアンカ文字列, および, ページ内の強調文字列に現れるかどうかによって決定する. 本ディレクトリは, 地域別に情報を表示する地域モードと, カテゴリ別に情報を表示するカテゴリモードの両方を提供する.

目 次

第 1 章

序論

近年, インターネットが急速に広まり, さまざまな情報が手軽に発信・受信できるようになった. これにより, ワールドワイドウェブ (WWW) は, 多くの分野を網羅する 1 つの大きな知識ベースを構成するに至っている. 地域情報も例外ではなく, 図 1.1に示すような自治体の公式ホームページを中心にさまざまな情報が充実してきている. それらの中には, 旅行に役立つ情報や自治体の行政サービスの案内など, 多くの有用な情報が含まれ, 旅行専門誌や広報誌に匹敵するほどの情報を WWW から得ることができる. また, WWW 上の地域情報は, 最近, 小学校で導入されつつあるインターネットを用いた教育の有用な教材として使われることも期待できる.

これらの地域情報を見つけ出す主要な方法には, サーチエンジンで検索する方法と, ディレクトリサービスやリンク集から辿る方法がある. しかし, これら 2 つの方法には, それぞれ問題点が存在する.

サーチエンジンによる検索には, 2 つの問題点がある. 1 つは不要なページが多く検索されるため, 求める情報を検索結果の中から探し出すことが必要となる点である. 特に, 地域情報をサーチエンジンで検索する際には, 「地域名」を入力文字列に選ぶことが多く, この文字列は多くのページに出現するため, 不要なページを検索しやすい. もう 1 つの問題点は, 特定のカテゴリに関する情報を複数地域にわたって調べるには不向きであるという点である. たとえば, 北陸地方の観光情報を WWW から探し出す場合, 北陸地方のそれぞれの地域に対するサイトから観光情報を探し出す作業を, 繰り返し行なう必要がある.

一方, ディレクトリサービスの問題点はデータの更新に手間がかかるという点である. そのため, 頻繁に更新することができず, 既に無いページをリンクしていたり, 最新の情報を含んでいないといった問題が生じる. 地域情報に特化したディレクトリサービスの 1 つである Cyber City Case Bank [1] は, 各地域の情報サイトへのリンクだけでなく, それら

のサイト内のページをカテゴリごとに分類しリンクしている, 非常に良くまとまったディレクトリサービスである. しかし, このサイトは 1998 年 7 月に更新を休止した. 休止理由は, 地域情報の普及という目的が達成されたためとしているが, 更新が容易ならば休止に至らなかったと考えられる.

これら 2 つの方法の問題点を解決するため, 本研究では地域情報ディレクトリを自動編集により生成するシステムを提案する. 自動化により, 大量の情報を高速に処理することが可能となり, WWW に存在する膨大な地域情報を処理することができる. また, 信頼できるサイトのみからの情報収集, 情報源となるサイトの各ページの分類, 目的に合わせた情報の提示などの編集処理により, ユーザは比較的簡単に求める地域情報を見つけ出すことができる.

本研究で作成した地域情報ディレクトリは, 日本全国の 47 都道府県と 3380 の市区町村の合計 3427 自治体に関する情報を提供する. このディレクトリは, 情報を地域ごとに表示する**地域モード**と, 複数の地域の特定のカテゴリに対するページを表示する**カテゴリモード**の 2 つの表示モードを持つ.

本システムでは, 次の 2 種類のサイトを情報源として利用する.

1. 地域サイト
   1 つの地域の幅広い情報を扱うサイト. 例えば, 図 1.1に示す, 自治体公式サイトなどがこれに該当する.

2. 特定情報サイト
   複数地域の特定の情報を扱うサイト. 例えば, 図 1.2に示す, 都道府県別面積サイトなどがこれに該当する.

本システムの主要部は, WWW 上のこれら 2 種類の情報源から地域情報を収集し, これらを編集する地域情報編集モジュールである. そこで行なう処理は, 上記の 2 種類の情報源によって異なる.

地域サイトを情報源とした地域情報編集処理は, 地域サイトの自動収集と, 地域情報ページの自動分類の 2 つである.

日本には 3427 の地域 (自治体) が存在する. これらの多くは, その地域に対する地域サイトを持つ. これらの地域サイトを全て手作業で見つけ出すことは, 非常に大変な作業になる. そこで, **地域サイトの自動収集**により, この作業を自動化する.

地域サイトに掲載されている情報は, 地域ごとにまとまっていて, 単に, 地域サイトの URL を見つけ出すだけでは, カテゴリモードを実現することはできない. カテゴリモード

図 1.1: 地域サイトの例 (http://www.city.uji.kyoto.jp/：京都府宇治市のサイト)

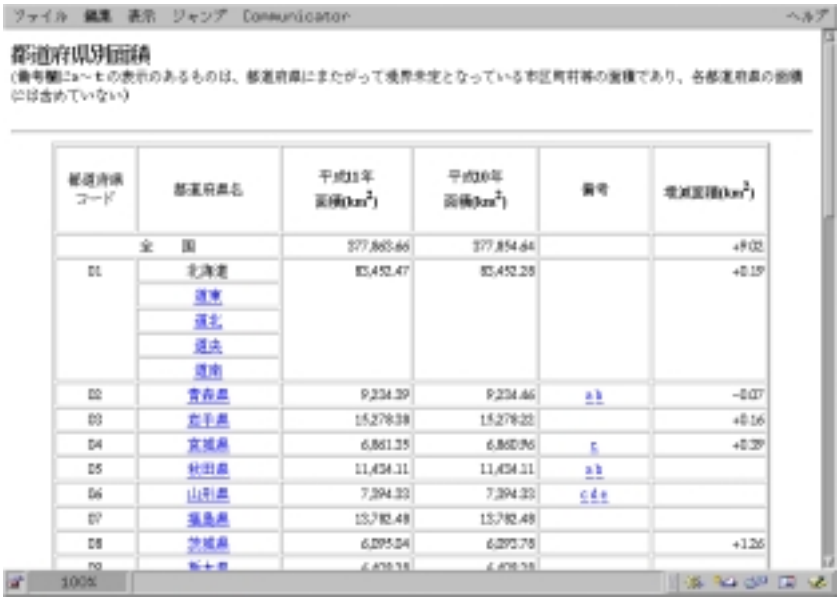

図 1.2: 特定情報サイトの例 (http://www.gsi-mc.go.jp/MAP/MENCHO/ichiran.htm：建設省国土地理院の都道府県別面積のサイト)

の実現には, 見つかったサイト内のページを, カテゴリごとに分類することが必要になる. **地域情報ページの自動分類**により, この処理を自動化する.

一方, 特定情報サイトの数はそれほど多くない. このため, このタイプのサイトの自動発見はそれほど重要ではない. これらのサイトにおいては, 情報は表形式で記述されていることが多い. **特定情報サイトからの情報抽出と情報生成**は, これらのサイトから, 表解析を用いて情報を抽出する. また, 抽出した情報を組み合わせて, 新しい情報を作り出す処理も行なう.

本論文では, まず, 第 2 章で地域情報ディレクトリとその自動生成の概要について説明する. 次に, 第 3 章で地域サイトの自動収集について説明し, 第 4 章で地域情報ページの自動分類ついて説明する. 第 5 章では, 特定情報サイトからの情報抽出と情報生成について説明する. 第 6 章では, システムを通しての検討と, 本研究と関連のある研究について述べる. 最後に, 第 7 章で本研究のまとめを述べる.

第 2 章

地域情報ディレクトリとその自動生成

本章では, 本研究で作成した地域情報ディレクトリとそれを実現するシステムの概要について説明する.

2.1 地域情報ディレクトリ

本研究で作成した地域情報ディレクトリは, 日本全国の 47 都道府県と 3380 の市区町村の合計 3427 自治体の情報を提供する. そのトップページを図 2.1に示す.

本ディレクトリは, 地域モードとカテゴリモードの 2 つの表示モードを持つ.

2.1.1 地域モード

地域モードは, 情報を地域ごとに表示するモードである. 図 2.2に, 地域モードのトップページを示す.

このページにおいて, 情報を表示させたい地域を指定する. 地域の指定は, ページ上部の地域名入力フォームに直接地域名を入力するか, あるいは, ページ下部に表示されている地域名から選択するかの, いずれかの方法を用いる.

地域名入力フォームに「辰口町」を入力すると, 図 2.3に示すページが表示される. ページ上部には, 辰口町の人口, 世帯数, 面積, 人口密度, 役所情報が表示される. ページ中程には, 自治体公式ホームページの候補サイトが表示され, ページ下部には, 辰口町の情報を掲載している自治体公式ホームページの候補サイト内のページが, 8 種類のカテゴリに分類されて表示されている. カテゴリ名のリンクを辿ると, 分類したページへのアンカがカテゴリごとに整理して表示される. たとえば, カテゴリ「観光・レジャー」のリンクを辿ると, 図 2.4に示すページが表示される.

図 2.1: 地域情報ディレクトリのトップページ

図 2.2: 地域モードのトップページ

図 2.3: 地域モードの情報表示例 (石川県 辰口町)

図 2.4: 辰口町の観光・レジャー

図 2.5: カテゴリモードのトップページ

2.1.2 カテゴリモード

カテゴリモードでは, 複数の地域 (地方) の, 特定のカテゴリに対するページを表示するモードである. このモードでは, 以下の 3 つの条件を入力し, 情報を選択する.

- 表示する情報のカテゴリ
- 情報をより詳細に限定するキーワード
- 対象地方

図 2.5に情報選択ページの例を示す. この図に表すように, カテゴリ「観光・レジャー」, キーワード「温泉」, 対象地方「北陸地方」を入力した場合, 図 2.6のページが表示される. このページは, 北陸地方のいずれかの地域の「観光・レジャー」に関するページの中で, キーワード「温泉」を含むすべてのページへのリンクが表示される.

2.2 地域情報ディレクトリ自動生成システムの概要

前節で示した地域情報ディレクトリを自動生成するシステムの全体構成を図 2.7に示す. 本システムは, 次の 3 つの要素から構成される.

図 2.6: カテゴリモードの情報表示例

1. **コンテンツデータベース**
   本地域情報ディレクトリで提供するすべての情報を格納したデータベース.

2. **表示モジュール**
   ユーザの要求に従って, コンテンツデータベースの内容を整形して表示するモジュール. 前節の地域モードとカテゴリモードの 2 つの表示モードを提供する.

3. **地域情報編集モジュール**
   本システムの主要部で, WWW 上の地域情報ページをもとに, 本ディレクトリで提供する地域情報を自動生成する. 生成した地域情報はコンテンツデータベースに格納される. このモジュールの構成を図 2.8に示す. このモジュールは, 地域サイトの自動収集, 地域情報ページの自動分類, 特定情報サイトからの情報抽出と情報生成の 3 つのサブモジュールから構成される. これらについては, それぞれ第 3 章, 第 4 章, 第 5 章で説明する.

以下では, まず, 2.3 節で本システムで用いる地域コード（地域を表現する形式）について説明する. 次に, 2.4 節でコンテンツデータベースについて説明する. 最後に, 2.5 節では表示モジュールについて説明する.

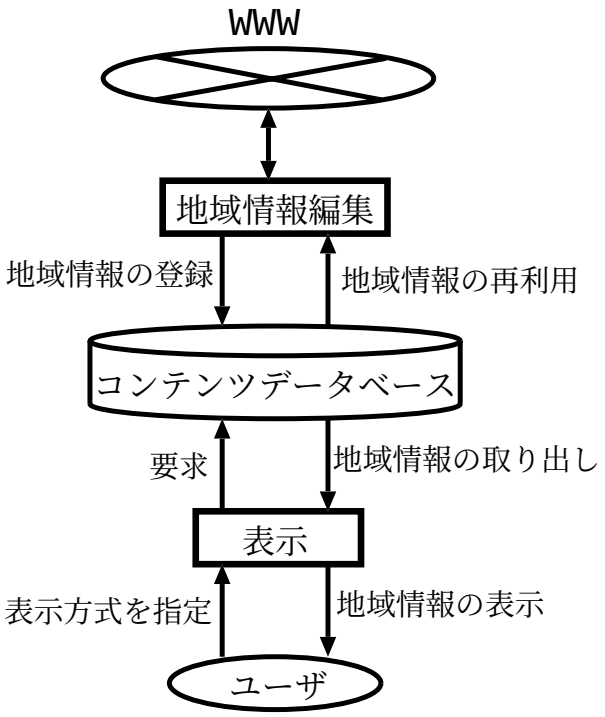

図 2.7: ディレクトリ自動生成システムの構成

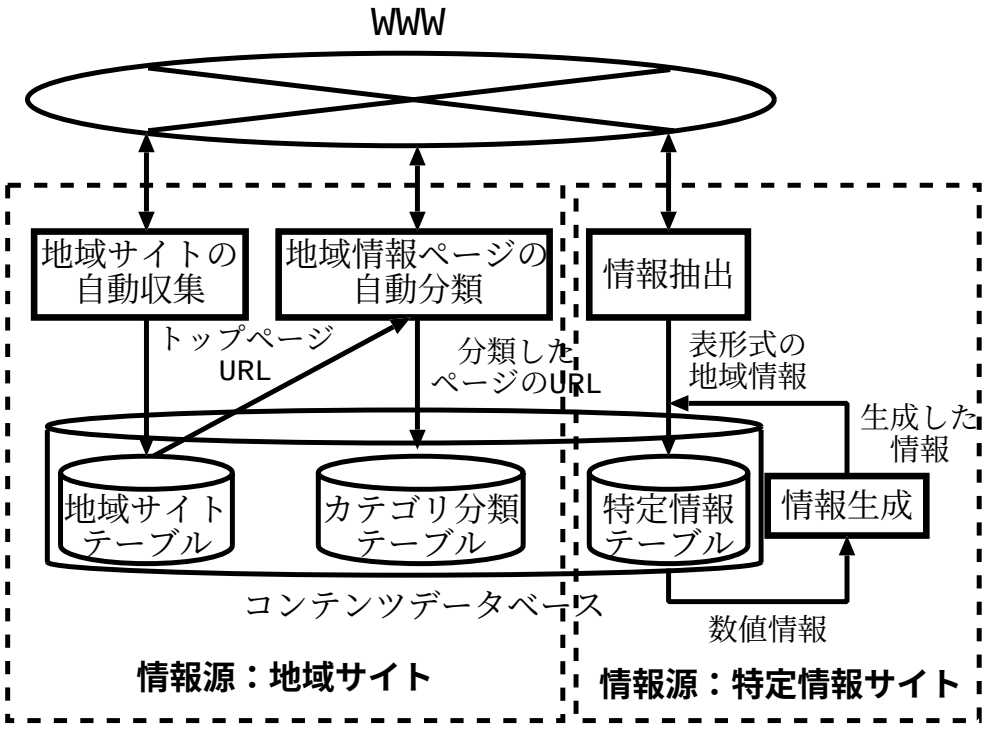

図 2.8: 地域情報編集モジュールの構成

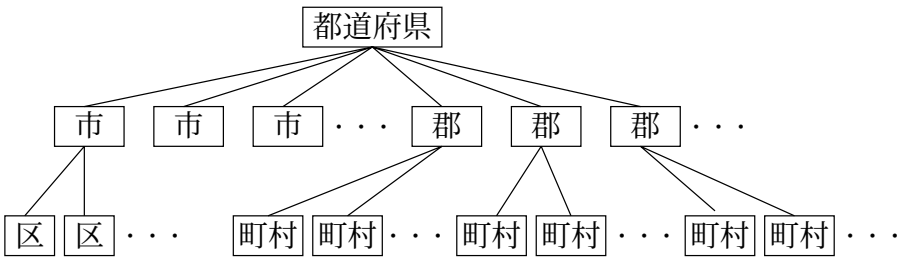

図 2.9: 地域レベルと包含関係

2.3 地域コード

2.3.1 地域コードの定義

本システムでは, 対象とする地域 (3427 自治体) と郡を表すために, **地域コード**と呼ぶ表現形式を用いる.

1 つの都道府県内の地域は, 図 2.9に示す構造をしている. ここでの, 都道府県, 市, 区, 郡, 町村を**地域レベル**と呼ぶ. それぞれの地域レベルには包含関係が存在する. 例えば, 郡レベルの下には, 必ず町村レベルの地域が存在し, 逆に, 町村レベルの 1 つ上のレベルは, 必ず郡レベルである.

地域コードは, 次の 3 つの条件を満たしていることが望ましい.

1. 対象とする地域と郡を一意に表すことができる.
2. コードのみから包含関係を容易に判定することができる.
3. コードのみから地域レベルを容易に知ることができる.

現在, 地域を表すコードには, 日本工業規格 (JIS) で定められた都道府県コード [2] と市区町村コード [3] がある. 都道府県コードは各都道府県に一意に割り当てられた数字 2 桁のコードであり, 市区町村コードは各都道府県内の市区町村に一意に割り当てられた数字 3 桁のコードである. つまり, 全国の市区町村は, 都道府県コードと市区町村コードを組み合わせた数字 5 桁のコードで, 一意に表現できる.

しかし, このコードを本システムの地域コードとして用いる場合, 次の 2 つの問題点がある.

1. 包含関係を知ることができない地域が存在する.
   市区町村コードは, 包含関係がわかるように設計されているが, 北海道や東京都の離

表 2.1: 地域コードの先頭文字と地域レベルの対応表

地域レベル	先頭文字
都道府県	A
市	B
区	C
郡	D
町村	E

島, 長崎県の対馬の町村の 1 つ上のレベルは, 本システムで採用している郡レベルでなく, 支庁レベルとして設計されている. このため, この地域の郡と町村の包含関係がわからない.

2. 地域レベルが簡単にわからない.
   例えば, ある数字 5 桁の JIS コードが, 区のレベルのコードであるかどうかを調べるためには, そのコードの 3 桁目が'1' で, かつ, 4 か 5 桁目が'0 以外の数字' となっているかどうかを調べる必要がある. このような複雑な条件は, 全てのレベルに存在し, 地域レベルを簡単に求めることはできない.

これらの問題を解決するために, 本研究では, この JIS コードを拡張した地域コードを用いる.

地域コードは, A〜E のアルファベット 1 文字と, それに続く 8 桁の数字から構成され, それぞれ以下の意味を持つ.

1. 先頭のアルファベット ··· その地域のレベルを表すアルファベット (表 2.1に各アルファベットが表す地域レベルを示す)
2. 最初の 2 桁の数字 ··· 都道府県コード
3. 次の 3 桁の数字 ··· 区, または, 町村の 1 レベル上の地域 (市, または, 郡) を同定するためのコード (表 2.2に示す. ただし, 支庁のある地域とは, 北海道内の郡, 東京都の離島, 長崎県の対馬である.)
4. 最後の 3 桁 ··· 市区町村コード

なお, 都道府県レベルの地域など, 市郡を表すコードや, 市区町村コードのない地域は, それらを意味する桁は 0 とする. 例えば, 石川県の地域コードは, 都道府県コードが 17 で, 市郡を表すコードと市区町村コードが存在しないので **A17000000** となる.

表 2.2: 1 レベル上の地域 (市, 郡) を同定するためのコード

地域レベル	コード
市区	市区町村コード
支庁のある地域の郡町村	同じ郡に含まれる町村の中で最も小さい市区町村コード
支庁のない地域の郡町村	同じ郡に含まれる町村の中で最も小さい市区町村コードの先頭 2 桁の後に 0 を付けた

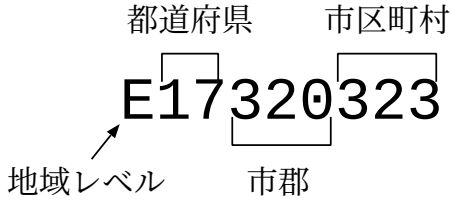

図 2.10: 石川県 能美郡 辰口町の地域コード

このような形式をとることで, 地域コードから, その地域のレベルや, 地域の包含関係が簡単にわかる. 例えば, 図 2.10に示すコードは, 石川県 辰口町のコードである. このコードの 1 つ上のレベルの地域コードを知りたい場合は, まず, このコードは町村 (E) のレベルなので, 上のレベルは郡 (D) となる. 次に, 郡レベルで存在する 1 桁目〜5 桁目のコードを, 辰口町の地域コードからそのままうつす (D17320). 最後に, 郡レベルでは存在しない市区町村コードが入る 6 桁目〜8 桁目に 0 を入れる (D17320000) と上のレベルの地域コードが完成する. これは能美郡のコードを表す. 逆に, 能美郡にある全ての地域の地域コードは, E17320... という形になっているため, 容易に求めることができる. このように, 地域コードのみから簡単に包含関係がわかるため, わざわざ地域名に変換して包含関係を調べたり, 包含関係を記述した対照表のようなものを用意しておく必要がない.

2.3.2 地域テーブルの作成

地域コードを作るための都道府県コードと市区町村コード, 全地域の地域名の基礎データとして, 郵政省のサイトにある郵便番号データファイル[1]を利用する. このデータファイルから地域コードと地域名の対応表である**地域テーブル**を作成する. 地域テーブルの 1 レコードは, 次の 5 つのフィールドから構成される.

[1] http://203.138.71.50/mobile/lzh/std/ken_all.lzh

表 2.3: 地域テーブルの例

地域コード	地域名 1(漢字)	地域名 2(漢字)	地域名 1(ひらがな)	地域名 2(ひらがな)
A17000000	石川県	石川県	いしかわけん	いしかわけん
B26204204	京都府 宇治市	宇治市	きょうとふ うじし	うじし
C12100105	千葉県 千葉市 緑区	緑区	ちばけん ちばし みどりく	みどりく
D29340340	奈良県 生駒郡	生駒郡	ならけん いこまぐん	いこまぐん
E25340343	滋賀県 野洲郡 野洲町	野洲町	しがけん やすぐん やすちょう	やすちょう

1. 地域コード

2. 地域名 1(漢字)
   都道府県名から記述した漢字地域名

3. 地域名 2(漢字)
   自治体名のみの漢字地域名

4. 地域名 1(ひらがな)
   都道府県名から記述したひらがな地域名

5. 地域名 2(ひらがな)
   自治体名のみのひらがな地域名

例として, 石川県 (地域コード：A17000000), 京都府宇治市 (地域コード：B26204204), 千葉市緑区 (地域コード：C12100105), 奈良県生駒郡 (地域コード：D29340340), 滋賀県野洲町 (地域コード：E25340343) のレコードを, 表 2.3に示す.

石川県の例で示すように, 都道府県レベルのレコードは, 地域名１と地域名 2 が同じものとなる. また, 地域名 1 では, 千葉市緑区や滋賀県野洲町の例のように, 属する市郡名も登録する. 郡は, 情報を提供する対象地域にはしていないが, 地域テーブルには, 全ての郡も登録されている.

地域テーブルは, 地域コードから地域名への変換や, 1 つ下のレベルの地域を求める時などに利用される.

表 2.4: 地域サイトテーブルの例

地域コード	URL(省略)	逆リンク数	IP アドレス	ページタイトル
A05000000	http://...	10217	210.136.179.1	美の国秋田ネット
A18000000	http://...	1898	202.239.99.41	?
B23100100	http://...	21541	202.232.59.29	名古屋市
B23100100	http://...	22	150.49.10.27	NAGOYA

2.4 コンテンツデータベース

コンテンツデータベースは, 地域情報ディレクトリで提供するすべての情報を格納するデータベースである. 本データベースは, **地域サイトテーブル**, **カテゴリ分類テーブル**, **特定情報テーブル**の 3 種類のテーブルから構成される.

2.4.1 地域サイトテーブル

地域サイトに関する情報を格納する. 地域サイトテーブルの 1 レコードは, 次の 5 つのフィールドから構成される.

1. 地域コード
   その地域サイトが対象としている地域

2. URL
   地域サイトのトップページの URL

3. 逆リンク数
   URL への逆リンク数[2]

4. IP アドレス
   地域サイトのサーバの IP アドレス

5. ページタイトル
   URL のページタイトル

地域テーブルの一部を表 2.4に示す.

[2]他のページからそのページへのリンク数.

表 2.5: カテゴリ分類テーブルの例 (一般)

地域コード	URL(省略)	文字列	特徴語
A32000000	http://...	■県章	県章
A36000000	http://...	徳島の紹介	徳島の紹介
B01203203	http://...	・小樽市のあゆみ	小樽市のあゆみ

福井県 (地域コード：A18000000) のレコードのように, URL の指すページにタイトルが記述されていない場合, ページタイトルには「?」を登録する. また, 名古屋 (地域コード：B23100100) のように, 1 地域に対して複数のレコードが登録されることもある.

地域サイトテーブルは, 地域サイトの自動収集で作成される. 地域サイトの自動収集は第 3 章で説明する.

2.4.2 カテゴリ分類テーブル

地域サイト内のページを 8 種類のカテゴリに分類したデータを保持する. カテゴリ分類テーブルには, それぞれのカテゴリに 1 つずつテーブルが存在し, 合計 8 つのテーブルから成る. これらのテーブルの 1 レコードは, 次の 4 つのフィールドから構成される.

1. 地域コード
   分類したページが対象としている地域

2. URL
   分類したページの URL

3. 文字列
   このページを分類する手掛りとなった文字列

4. 特徴語
   文字列の中に含まれている特徴語

カテゴリ「一般」のカテゴリ分類テーブルの一部を表 2.5に示す.

カテゴリ分類テーブルは, 地域情報ページの自動分類で作成される. 地域情報ページの自動分類は第 4 章で説明する.

2.4.3 特定情報テーブル

特定情報サイトにある表のフィールド名を属性, その値を属性値として格納する. 特定情報テーブルは, 数値情報用テーブルと, テキスト情報用テーブルの 2 つから成り, 属性値が数値である場合は数値情報用テーブルに格納し, テキストである場合はテキスト用テーブルに格納する.

数値情報用テーブルは, 次の 8 つのフィールドから構成される.

1. 地域コード
   数値情報が対象としている地域

2. URL
   情報源となるページの URL

3. 属性
   情報源となる表のフィールド名

4. 属性値
   数値情報の値

5. 全国順位
   全国での順位

6. 全国の地域総数
   このラベルの数値情報が存在する全国の地域の総数

7. 都道府県内順位
   同じ都道府県内での順位

8. 都道府県内の地域総数
   このラベルの数値情報が存在する同じ都道府県内の地域の総数

例として, 石川県辰口町 (地域コード：E17320323) の全レコードを表 2.6に示す.
一方, テキスト情報用テーブルは, 次の 4 つのフィールドから構成される.

1. 地域コード
   テキスト情報が対象としている地域

表 2.6: 特定情報テーブル (数値情報テーブル)

地域コード	URL(省略)	属性	属性値	全国順位	全国の地域総数	都道府県内順位	都道府県内の地域総数
E17320323	http://...	市区町村番号	17323				
E17320323	http://...	団体コード	173231				
E17320323	http://...	人口	13113	1347	3192	16	41
E17320323	http://...	世帯数	3874	1356	3192	17	41
E17320323	http://...	平成 10 年面積 (km2)	57.13	1700	2967	22	39
E17320323		人口密度	229	1303	2939	19	39

表 2.7: 特定情報テーブル (テキスト情報テーブル)

地域コード	URL(省略)	属性	属性値
E04300301	http://...	平成 10 年面積 (km2)	境界未定
E04300301	http://...	電話番号	0224-33-2211
E04300301	http://...	新郵便番号	989-0821
E04300301	http://...	ふりがな	ざおうまち
E04300301	http://...	住所	刈田郡蔵王町大字円田字西浦北１０

2. URL
   情報源となるページの URL

3. 属性
   情報源となる表のフィールド名

4. 属性値
   テキスト情報の内容

宮城県蔵王町 (地域コード：E04300301) の全レコードを表 2.7に示す.

特定情報テーブルは, 特定情報サイトからの情報抽出と情報生成で作成される. 特定情報サイトからの情報抽出と情報生成は第 5 章で説明する.

表 2.8: 地域情報の検索テーブルと表示場所

地域情報の項目	テーブル	表示場所
人口, 世帯数, 面積, 人口密度, 役所情報	特定情報テーブル	ページ上部
自治体公式ホームページの候補	地域サイトテーブル	ページ中程
分類した地域サイト内のページ	カテゴリ分類テーブル	ページ下部

2.5 表示モジュール

表示モジュールは, ユーザの要求に合ったディレクトリを, コンテンツデータベースにある地域情報から生成し提供する. このシステムは, 地域モードとカテゴリモードの 2 種類の表示モードを持つ.

2.5.1 地域モード

地域モードは, 情報を地域ごとに表示する. 表示モジュールは地域モードのトップページで地域名 (文字列) が入力されると, 地域テーブルを参照して, 入力された地域に対応する地域コード求め, それを用いてコンテンツデータベースから必要な情報を検索し, それらの情報を整形して表示する.

例として, 石川県辰口町の地域モードの地域情報表示ページを図 2.3に示す.

トップページで「辰口町」が入力されると, まず, 地域テーブルから辰口町に対応する地域コード (E17320323) を取り出す. その地域コードを用いて, それぞれのテーブルから, 表 2.8に示す地域情報を検索し, それぞれ決められた場所に整形して表示する.

2.5.2 カテゴリモード

カテゴリモードは複数地域 (地方) の特定のカテゴリに対するページを表示する. 表示モジュールは, カテゴリモードのトップページで, カテゴリ, キーワード, 対象地方が選択されると, 地域テーブルから選択された地方にある全ての地域の地域コードを取り出し, それらを用いて指定されたカテゴリに対応するカテゴリ分類テーブルから, それらの地域コードで格納されている情報を検索し, それらの情報を整形して表示する.

カテゴリモードの情報表示ページの例を図 2.6に示す. このページは, トップページでカテゴリ「観光・レジャー」, キーワード「温泉」, 対象地方「北陸地方」とした場合に表示する. この場合, 「観光・レジャー」に対応するカテゴリ分類テーブルに格納しているレ

コードの中から, 北陸地方にある地域に対応する地域コードで, 文字列の中に「温泉」を含んでいるものを全て検索し, 整形して表示する.

第 3 章

地域サイトの自動収集

本章では, 地域サイトを情報源とした最初の編集処理である, 地域サイトの自動収集について説明する. このシステムが対象としている地域数は, 全部で 3427 地域あり, 多くの地域がその地域を対象とした地域サイトを持つ. これらのサイトを, すべて, 手作業で見つけ出すことは非常に困難な作業であるため, 地域サイトの自動収集を行なう. 地域サイトの自動収集では, WWW から地域サイトのトップページの URL を探し出し, コンテンツデータベースの地域サイトテーブルに登録する処理を行なう.

3.1 URL パターンとリンク集を利用した地域サイトの収集

WWW 上には, 図 3.1で示すような, 地域サイトのリンク集が多数存在する. これらのリンク集には, 複数の地域サイトの URL が列挙されているので, このようなリンク集を見つけることができれば, そのページから地域サイト URL を容易に収集することができる. 地域サイトの自動収集では, このようなリンク集を, 地域サイトに見られる一般的な URL のパターンをもとに, 既存のサーチエンジンで発見し, このリンク集を用いて, 地域サイトのトップページの URL を収集する.

これにより, 大量の URL が自動的に収集できるが, 収集した URL の中には, 対象ページの存在しない URL(ダングリングリンク [5]) や, 地域サイトのトップページ以外のページの URL も含まれる. ここでは, このような URL に対しても適切な処理を行なう必要がある.

以下では, まず, 3.2 節で地域サイトリンク集の発見方法について説明する. 次に, 3.3 節で発見した地域サイトリンク集から地域サイトの URL と, それに対応する地域の抽出方法について説明する. 3.4 節ではダングリングリンクの処理について説明し, 3.5 節では

図 3.1: 地域サイトリンク集の例 (http://www.nippon-net.ne.jp/miyazaki.html：Nippon-net 宮崎県)

トップページ以外のページに対応する URL の処理について説明する. 最後に, 3.6 節でこの章で説明する方法での地域サイトの自動収集の実験結果と検討を述べる.

3.2 地域サイトリンク集の発見

本節では, 地域サイトのトップページ URL の抽出元となる, 地域サイトリンク集を自動発見する方法を説明する. 地域サイトリンク集の数は, 地域サイトの数ほど多くはないが, これらを手作業で見つけ出すことは大変な努力が必要となる. このため, 自動的に, 多くの地域サイトリンク集を発見できる方法が求められる.

ここでは地域サイトのトップページ URL である可能性が高い URL 群を用いて, サーチエンジンを用いて地域サイトリンク集を検索する. 地域サイトのトップページ URL の可能性が高い URL は, 地域サイトリンク集を発見するための “種” として使われるため, ここではこれらの URL を「種 URL」と呼ぶことにする.

また, 地域サイトリンク集の 1 ページには, あるグループ (地域グループ) ごとに掲載さ

れていることが多い. このため, サーチエンジンで地域サイトリンク集を検索する際にも, この地域グループごとに検索処理を行なう.

3.2.1 種 URL

地域サイトリンク集を検索するために “種” として使われる, 地域サイトのトップページ URL の可能性が高い URL を「種 URL」と呼ぶ. ここではまず, 次のパターンを持つ URL が, しばしば地域サイト URL として用いられることに着目する.

http://www. **地方公共団体ドメイン名**/

ここでの, 「地方公共団体ドメイン名」は以下のルールで作成されるドメイン名を指している.

属性. **地域名**. **都道府県名**.jp

「属性」には, それぞれの地域に応じて, pref (都道府県), city (市・特別区), town (町), vill (村) を代入する. また, 「地域名」と「都道府県名」には, その地域の地名をローマ字で代入する. なお, 地方公共団体が都道府県の場合は, 地域名の部分は存在しない. また, 政令指定都市の場合は, 「都道府県名」を省略する. 本ドメイン名は, 日本ネットワークインフォメーションセンター (JPNIC) が定義する, 地方公共団体のためのドメイン名 [4] である. たとえば, 石川県金沢市のドメイン名は **city.kanazawa.ishikawa.jp** であり, それに対応する URL は次のようになる.

http://www.city.kanazawa.ishikawa.jp/

全ての地域サイトの URL がこのパターンの URL となっているわけではないが, このパターンの URL を持つサイトが, 求める地域サイトであることはほぼ間違いない. つまり, このパターンの URL を種 URL として用いることができる.

種 URL を多数リンクしているページは, 求める地域サイトリンク集である可能性が高い. このようなページは, サーチエンジン (Infoseek[1]) の「リンク検索機能」を用いて検索することができる. 「リンク検索機能」とは, クエリとして入力した URL をリンクしているページを検索する機能である. ここでは, 複数の種 URL をクエリとして, リンク検索機能を用いて検索する.

[1] http://www.infoseek.co.jp/

ただし, Infoseek では, 検索の条件が 10 個以上だと結果が得られないため, 逆リンク数の多い URL を 10 個を選んで入力する. 逆リンク数の多い URL を選ぶのは, 以下の 2 つの理由からである.

- 逆リンク数の多い URL はリンク集に掲載されている可能性が高い.
- 逆リンク数の多いページは, 他の多くのページから支持されているということである. それは, 信頼性が高く, 情報が充実していることを表している. 支持されているページを多くリンクしているリンク集に掲載されるリンク先のページは, 同様に, 信頼性が高く, 情報が充実している可能性が高いと考えられる.

地域サイトの自動収集処理は, 新たな地域サイトの URL を収集するために繰り返し実行する. このため, 地域サイトリンク集の自動発見も繰り返し行なう必要がある. 初回の処理での種 URL は, すべて, 地方公共団体ドメイン名に対応する URL である. しかし, すべての有用な地域サイトが, 地方公共団体ドメイン名に対応する URL を使っているわけではなく, より多くの地域サイトリンク集を見つけるためには, 地方公共団体ドメイン名に対応する URL ではなくても, 他のページから支持されているページの URL を種 URL とする必要がある.

このことから, 2 回目以降の収集処理では, その時点で地域サイトテーブルに登録されている URL で, 逆リンク数の多いものを種 URL とする. ただし, 地域サイトリンク集の多くは, 1 地域につき 1 つだけ URL を掲載しているため, 種 URL は 1 地域につき逆リンク数の多い URL を 1 つだけ出す. これにより, 新たな地域サイトリンク集が発見でき, 多くの地域サイト URL を収集することができる.

3.2.2 地域グループ

地域サイトリンク集の 1 ページには, 47 都道府県のみか, 同じ都道府県内に存在する市町村, 同じ政令指定都市に存在する区ごとにリンクが掲載されている場合が多い. 例えば, 図 3.1に示すページは, 宮崎県にある地域のみをリンクしている. 多くの地域サイトリンク集を発見するため, まず, 表 3.1に示す 3 種類の地域グループを設定する.

表 3.1の, グループ名が「全国」のグループの構成要素は, 北海道, 青森県, 秋田県, …, 鹿児島県, 沖縄県の 47 都道府県である. また, グループ名が「都道府県」のグループの構成要素は, 同じ都道府県内の市町村で, 例えば北海道なら, 札幌市, 旭川市, 函館市, 長万部町, 美瑛町, 白滝村など, 北海道にある全ての市町村が 1 つのグループとなる. 「都道府県」グループは, 都道府県ごとに 1 つのグループができるため, 全部で 47 グループある. グルー

表 3.1: 地域サイトリンク集検索のためのグループ

グループ名	構成要素	グループ数
全国	47 都道府県	1
都道府県	同じ都道府県内の市町村	47
政令指定都市	同じ政令指定都市にある区	12

プ名が「政令指定都市」のグループの構成要素は, 同じ政令指定都市にある区で, 例えば, 札幌市なら, 中央区, 北区, 東区, 白石区, 豊平区, 南区, 西区, 厚別区, 手稲区, 清田区の, 札幌市にある全ての区が 1 つのグループとなる. 「政令指定都市」グループは, 政令指定都市ごとに 1 つのグループができるため, 2000 年 2 月現在, 全部で 12 グループある.

地域サイトリンク集の検索処理は, 種 URL をこれらのグループに分けて, それぞれのグループごとに行なう. これにより, 多くの地域サイトリンク集を発見することができる.

3.2.3 アルゴリズム

以上の考え方に基づき, 次のアルゴリズムで地域サイトリンク集を収集する.

1. 地方公共団体ドメイン名に対応する URL を, 全ての地方公共団体に対して作成する. そのうち, 実際にページが存在する URL をコンテンツデータベースの地域サイトテーブルに登録する.

2. 表 3.1に示す「全国」「都道府県」「政令指定都市」の 3 種類のグループを設定し, それぞれのグループに対して, 以下の処理を行なう.

   (a) 種 URL として, そのグループに属する全ての地域に対応する地域サイトの URL を, 地域サイトテーブルから, 1 つの地域に対して 1 つだけ取り出す. 1 つの地域に対して複数の地域サイト URL が登録されている場合は, 逆リンク数の最も多いものを選ぶ.

   (b) 取り出した種 URL 群の中から, 逆リンク数が多い URL を上位 10 個選ぶ. これらの URL をできるだけ多くリンクしているページをサーチエンジンを用いて検索する.

こうして見つかったページを地域サイトリンク集の候補とする.

3.3 地域サイト URL の抽出

地域サイトリンク集の候補を発見すると, そのページから地域サイトの URL と, その URL が対象としている地域名を抽出する.

3.3.1 抽出方法

地域サイトの URL は, 地域サイトリンク集の候補にあるページのアンカから抽出し, その地域サイトの URL が対象としている地域名は, その URL を抽出したアンカのアンカ文字列[2]から抽出する. しかし, 地域サイトリンク集内の, アンカ文字列に地域名を含むアンカのすべてから URL を収集したのでは, 有用でないサイト (例えば, 社団法人京都府建築士事務所協会など) も数多く収集してしまう.

図 3.1にも示すように, 地域サイトリンク集のアンカ文字列は地域名のみの場合が多い. そこで, アンカ文字列が厳密に地域名であるアンカのみから URL を抽出する.

ここで問題となるのは, 地域名が同じ地域が全国に数多く存在する点である. 例えば, 群馬県内だけでも「東村」が勢多郡, 吾妻郡, 佐波郡のそれぞれに存在し, 全国では, 5 つの「東村」が存在する. アンカ文字列に郡名が記述されている場合には問題は生じない. しかし, 実際は「東村」のような自治体名のみをアンカ文字列としている場合がほとんどで, アンカ文字列だけからは, 地域を特定することができない.

そこで, ほとんどの地域サイトリンク集が都道府県ごとに掲載していたり, 町村を郡ごとに整列して掲載していることに着目する. つまり, URL を抽出するアンカの前に記述されているアンカは同じ都道府県で, かつ, 同じ郡の地域に対するアンカである可能性が高く, これを利用する. 実際には, 同名地域の存在する地域名がアンカ文字列となっていた場合は, それらの全ての地域の地域コードと, 直前に抽出した URL に対応する地域の地域コードを 1 桁目から照合し, 1 桁目から連続して一致した桁数が最も多い地域コードを持つ地域がこのアンカに対応する地域であると判断する.

3.3.2 アルゴリズム

以上の考え方に基づき, 次のアルゴリズムで地域サイトの URL, およびその URL が対象としている地域を収集し, コンテンツデータベースの地域サイトテーブルに登録する.

1. 発見した地域サイトリンク集から, アンカ文字列が地域名となっているアンカを見つけ, その URL を抽出する. 同名の地域名が存在する場合は, 直前に抽出した URL

[2]アンカタグ (<a>..) で囲まれる文字列.

の対象とする地域コードと照合し地域を特定する.

2. 抽出した URL の逆リンク数, サーバの IP アドレス, 対応するページのタイトル, 対象とする地域の地域コードを調べる.

3. 地域コード, 抽出した URL, 逆リンク数, IP アドレス, ページタイトルを 1 レコードとして, 地域サイトテーブルのそれぞれ適切なフィールドに登録する.

3.4 ダングリングリンクの処理

ここでは, 石田ら [5] がダングリングリンクと呼ぶ, 切断リンクの処理について説明する. ダングリングリンクとなっている URL には, 対応するページが WWW 上のどこにも存在しない, もしくは, 存在していたとしても何処にあるのかが全くわからない URL と, 対応するページが移動していて, 移動先がわかる URL の 2 種類がある.

前者の URL は不必要であるため, 地域サイトテーブルに登録しない.

後者は, 移動先のページが地域サイトのトップページである可能性が高いため, 移動先の URL を探し出し, 元の URL の代わりに地域サイトテーブルに登録する処理を行なう. この処理は, 「移動通知ページ」(図 3.2, 図 3.3に例を示す) と「ne ドメイン変更した or ドメイン」に対して行なう. 以下にそれぞれの処理方法を説明する.

3.4.1 移動通知ページ

移動通知ページには, 移動先の URL が明示的に書かれている場合が多く, これを抽出することは容易である. そのため, 移動通知ページが判定できれば簡単に移動先の URL を見つけられる. ここでは, これらのページに見られる, 以下のような特徴を用いて, 移動通知ページの判定を行なう.

- 図 3.2に示すように, 「引越」や「移動」など, 移動を示すキーワードが書かれていることが多い.

- 図 3.2, 図 3.3に示すように, 移動先の URL 自体がアンカ文字列となっているアンカが存在する.

実際には, 文献 [5] のリダイレクションに対する処理を参考にした, 次の方法で行なう.

まず, 登録しようとしている URL のページが, 以下の 2 つの条件をいずれも満たした時は, そのページが移動通知ページであると判定する.

表 3.2: 移動を示すキーワード (正規表現)

1. 知らせ
2. 移動
3. 移転
4. 変更
5. リニューアル
6. 変わ
7. 引越
8. 引っ越
9. ジャンプ
10. 下記 (に | へ | を)
11. 以下 (に | へ | を)
12. (新しい | 下記の | 以下の | 次の)(アドレス | ホームページ |URL)
13. 自動的に
14. (アドレス |URL) が
15. こちら

- ページのサイズが 2KBytes 以下 (文献 [5] では, 4KBytes 以下となっている).
- 表 3.2にある 15 パターンの移動を示すキーワードのうち, 2 種類以上含む.

この条件で対象としている, 移動通知ページの例を図 3.2に示す.

また, 移動先の URL 自体をアンカ文字列とした移動通知ページが多く存在することから, 次の 2 つの条件を満たすページも移動通知ページであると判定する.

- ページのサイズが 500Bytes 以下.
- ページ内のアンカが 2 つ以下で, アンカ文字列と移動先の URL が一致するアンカが含まれる.

この条件で対象としている, 移動通知ページの例を図 3.3に示す.

文献 [5] では, HTTP のリダイレクション機能[3]などによる自動ジャンプは, これらの条件を満たさなくても移動通知ページと判定しているが, 地域サイトの中には, 北海道の公式サイト[4]のように, リダイレクションを持たせたページを表紙として使っている場合があるため, 本研究では上記の条件のみで移動通知ページを判定する.

[3] <head>..</head>の中に<meta http-equiv="refresh" content="秒数;url=移動先 URL">を記述

[4] http://www.pref.hokkaido.jp/

図 3.2: 移動通知ページの例 1(http://www.tsc.co.jp/ĝama/：愛知県蒲郡市)

図 3.3: 移動通知ページの例 2(http://www.town.tatsunokuchi.ishikawa.jp/：石川県辰口町)

上記の条件を満たし, 移動通知ページであると判断されると, 単純に, そのページから 1 つ URL を取り出し, その URL を元の URL に代わって地域サイトテーブルに登録する.

3.4.2 ne ドメインに変更した or ドメイン

以前, or ドメインを使用していた ISP[5] の URL は, ne ドメインに変更されたものが多いため, 対応するページが存在しない URL で or ドメインのものは, or を ne に変更されている可能性がある. そこで, or ドメインの URL に対する不要 URL の処理は, ne ドメインの URL に対応するページが存在するかを確認し, 存在した場合は ne ドメインの URL を地域サイトテーブルに登録する.

実際には, 次の方法で行なう.

まず, 登録する URL が次の条件を全て満たす時, ne ドメインに変更した or ドメインのダングリングリンクと判定する.

- URL が or ドメインである.
- URL にページが存在しないか, 文書内にその URL の'or' を'ne' に変換した, ne ドメインの URL の記述がある.
- ne ドメインの URL にページが存在する.

[5]Internet Service Provider の略. 一般にプロバイダと呼ばれる.

上記の条件を満たしたとき, 登録しようとしていた or ドメインの URL に代わって, ne ドメインの URL を地域サイトテーブルに登録する.

3.5 トップページ以外の URL の処理

トップページ以外のページは, その地域の議会のページや, 統計のページなど, 専門的なトピックのページである場合が多く, それらのページのほとんどが, そのサイトのトップページから簡単に辿れる. これらのページは, ユーザに多くの公式サイトの候補を提示して, 混乱させる原因になったり, この次に行なうページの分類で, 余分な処理をしてしまう原因になる. このため, これらのページの URL は地域サイトテーブルには登録しない.

ほとんどの地域サイトでは, サイト内のページの中で, 最も逆リンク数の多いページが, サイトのトップページである. これを利用して, そのサイト内で最も逆リンク数の多いページの URL のみを地域サイトテーブルに登録する方法で, トップページ以外の URL は削除できる. ただし, 同じサイトで違う地域のサイトが存在する場合もあるため, ここでの処理は, 同じ地域でサイトのトップページとして収集した同一サイトのページは, 最も逆リンク数の多いページの URL を 1 つだけ登録する.

同一サイトのページとは, 同一のサーバから発信されたページのことを表す. 複数のページが, 同一のサーバから発信されたかどうかの判定の, 最も簡単な方法は, URL のサーバ名が同じかどうかを調べることである.

しかし, 1 つのサーバには, 複数のサーバ名を付けることができる. 例えば, 石川県金沢市の地域サイトの URL である以下の 2 つのは, 異なるサーバ名だが, 同一サーバから発信されたページで, これらの URL に対応するページの内容も全く同じである.

http://www.city.kanazawa.ishikawa.jp/
http://city.kanazawa.ishikawa.jp/

このことから, サーバ名のみで, 同一サーバから発信されているページかどうか判定する方法は, 不十分である. そこで, URL のサーバ名からサーバの IP アドレスを調べ, IP アドレスが同一のサーバ名を含む URL のページは同一のサーバから発信されたページと判定する.

既に地域サイトテーブルに登録している URL と同一のサーバから発信されたと判定された URL を登録しようとしたときを考える. 既に登録されている URL に対応するページの方が逆リンク数が多い場合は, 新たに登録しようとしている同一サイトの URL は登録しない. 逆に, 新たに登録しようとしている URL に対応するページの方が逆リンク数が多い場合は, 既に登録している同一サイトの URL を消去して新たに URL を登録する.

表 3.3: 地域サイトの収集数の実験結果

	カバー地域数	収集サイト数
URL 作成時	719(21.0%)	719(1.00)
1 回目終了時	2725(79.5%)	3532(1.30)
2 回目終了時	2852(83.2%)	4012(1.41)

表 3.4: 有用・無用地域サイト数の実験結果

作成者	有用なサイト	無用なサイト
該当地域の役所	39	5
上位レベルの役所	4	0
役所以外の公共団体	14	0
企業	9	1
個人・作者不明	7	3
合計	73	9

3.6 実験と検討

3.6.1 実験

上記の方法に基づき, 地域サイトを収集する実験を行なった. 地域サイトの収集数の実験結果を表 3.3に示す.

この表の URL 作成時とは, 地方公共団体ドメイン名に対応する URL を用いた場合を示す. このパターンの URL によって見つけることのできる地域は 719 地域で, 全体の 21%であった. この URL 群を用いて, 地域サイトのリンク集を収集し, 地域サイトの URL を収集した結果, 2725 地域, 全体の 79.5%の地域に対して地域サイトを見つけることができた. これをもう 1 度繰り返したところ, さらに 127 地域に対して地域サイトを見つけることができ, 最終的に 2852 地域, 全体の 83.2%の地域に対して地域サイトが見つけることができた.

収集したサイトの総数は, 1 回目終了時で 3532 サイトであり, 1 地域当りの平均サイト数は 1.30 であった. 2 回目終了時には 4012 サイトで 1 地域当りの平均サイト数は 1.41 であった.

また, 収集した地域サイトが有用か無用かの実験を, 全国 56 の地域, 82 サイトを対象に

行なった. 実験結果を表 3.4に示す. この実験での有用なサイトとは, その地域に関する情報が掲載されていて, かつ, このあと行なうページ分類処理で, 情報源となりうるサイトとした.

該当地域の役所が作成者のサイトには,「掲載情報に関する問い合わせ先」が役所になっているサイトや, 公式サイトと明記しているサイトなどを分類した. また, 上位レベルの役所とは, 該当地域が属する都道府県や支庁を表している. 役所以外の公共団体は, 例えば, 商工会議所や観光協会, 教育委員会などを指している.

該当地域の役所が作成したサイトで, 無用と判定した 5 サイトの内, 3 サイトは地方公共団体ドメイン名を含む URL を持つサイトで, その全てが,「ただいま作成中」と記述しており, 今後, 有用な地域サイトとなりうるサイトであった. また, 残りの 2 サイトは, 有用なサイトと判定したサイトと全く同じ内容のサイトであった. 作成者が個人・作者不明の 3 つの無用サイトは, 情報提供が中止されたサイトで, 以前は該当地域の地域サイトとして情報を提供していたようであった. この実験の対象サイトで, 該当地域と全く関係のないサイトだったのは, 作成者が企業の無用なサイトのみであった.

全 82 サイトの内, 約 90%の 73 サイトが有用なサイトと判定できるサイトであり, 無用なサイトの中でも, 全く該当地域と関係のないサイトはたった 1 つであった.

3.6.2 検討

表 3.3に示す地域サイトの収集数の実験結果より, この章で説明した方法で大量の地域サイトを収集できることがわかった.

また, 表 3.4に示す, 地域サイトの有用・無用調査実験の結果から, この方法により収集した地域サイトの, ほとんどが有用なサイトであり, その多くが, 該当地域の役所が作成した信頼性の高く, 内容も豊富なサイトであるといえる.

以上のことから, 本章で提案した, URL パターンと地域サイトリンク集を利用した地域サイトの収集は, 非常に有効な収集方法であるといえる.

地域サイトの収集数の実験結果で示すように, 現在, 地方公共団体ドメイン名に対応する URL は, 全地域数の 21%の地域でしか使われていない. しかし, 地域サイトの有用・無用調査の実験結果で, 該当地域の役所が作成した無用なサイトと判断された 5 サイトのうち 3 サイトが地方公共団体ドメイン名に対応する URL であり,「ただいま作成中」のサイトであることから, 地方公共団体ドメイン名に対応する URL は, 今後, 多くの地域の公式サイトの URL として使われることが予想できる. 実際, 8ヶ月前に行なった地域サイトの収集数の実験では, 地方公共団体ドメイン名に対応する URL は, 全地域の約 17%の地

域でしか使われていなかったが, 現在は 21%にまで増えている.

本章で提案した方法では, 地方公共団体ドメイン名に対応する URL を持つ地域サイトは, サーチエンジンや地域サイトリンク集に依存しないで見つけ出すことができる. つまり, 頻繁に更新を行なえば, 新たに開設された地方公共団体ドメイン名に対応する URL を持つ地域サイトを, サーチエンジンや地域サイトリンク集よりも早く発見し, 本ディレクトリの情報源とすることができる. このような点からも, この方法は有効な収集方法であるといえる.

第 4 章

地域情報ページの自動分類

本章では, 前章で説明した地域サイトの自動収集により収集した, 地域サイトにあるページをカテゴリに分類する, 地域情報ページの自動分類について説明する.

地域サイトの情報は, 地域ごとにまとまっていて, 単に, 地域サイトの URL を見つけ出すだけでは, カテゴリモードを実現することはできない. このためには, 見つかったサイト内のページをカテゴリごとに分類し, 組織化することが必要になる. ここでは, 地域サイト内のページを, 8 つのカテゴリに自動的に分類する処理を行ない, カテゴリモードの実現を可能にする. 分類したページの URL は, コンテンツデータベースのカテゴリ分類テーブルに登録する.

4.1 特徴的表現を利用したページの自動分類

地域情報ページの自動分類では, 地域サイトの自動収集により収集した地域サイトのトップページ URL を入力とし, 地域サイトにあるページを 8 つのカテゴリに分類する. この処理は, カテゴリモードを実現するために必要となる.

それぞれのページをどのカテゴリに分類するかは, 次の 3 つの文字列に, カテゴリに特有の単語や表現 (特徴語) が現れるかによって決定する.

- 他のページにある, 分類するページをリンクしているアンカの**アンカ文字列**
- 分類するページの**タイトル**
- 見出しや強調文字など, 分類するページ内の目立つ文字列 (以後, **強調文字列**と呼ぶ)

これらの文字列を**判定文字列**と呼ぶ. それぞれのカテゴリに特有の単語や表現である特徴語は, 特徴語辞書としてまとめておき, ページの持つ判定文字列にカテゴリの特徴語が含まれる場合, そのカテゴリにそのページを分類する.

もし, 判定文字列の中に複数カテゴリの特徴語が含まれている場合は, そのページに複数カテゴリにわたる内容が掲載されている可能性が高いため, 含まれている特徴語に対応する全てのカテゴリに分類する.

地域サイト内の全てのページを分類すると, 1 つのカテゴリに多くのページが分類されてしまうことがある. 表示モジュールでこれらの全てをユーザに表示してしまうと, ユーザがその中から知りたい情報を探し出す手間が余計にかかってしまう. 同じカテゴリに分類される多くのページは, その中の幾つかのページから容易に辿れ, そのようなページのみを分類するほうが簡単に知りたい情報を探し出せる場合が多い. ここでは, 無駄なページを分類しないために, ページのカテゴリ分類処理の対象とするページを限定する.

以下では, まず, 4.2 節で分類カテゴリについて説明する. 次に, 4.3 節で特徴語辞書について説明する. 4.4 節では, 分類対象ページを説明し, 4.5 節で判定文字列について説明する. 4.6 節では, 自動分類のアルゴリズムを説明し, 最後に 4.7 節でこの章で説明する方法でのページ分類の実験結果と検討を述べる.

4.2 分類カテゴリ

分類カテゴリとして, 次の 8 つのカテゴリを用いる.

1. **一般**
   地勢や首長のあいさつ, 姉妹都市, 地域の木など, 一般的な事柄

2. **計画・産業**
   産業に関する情報と, 街づくり, 企業誘致などの計画に関する情報

3. **イベント・祭り**
   祭りや行事, 公演などのイベント情報

4. **文化・歴史・教育**
   文学, 伝統文化, 芸術などの文化と, 史跡, 地域の歩みなどの歴史, および講習や教育に関する情報

5. **観光・レジャー**
   観光名所や宿泊所, 名産品などの観光情報と, キャンプ場やスポーツ施設などのレジャー情報

6. **統計**
   人口, 面積など各種の統計情報

7. **住民向け**
   災害情報や, 役所のサービスなど住民向けの情報

8. **リンク**
   地域と関連あるサイトへのリンク集

分類するカテゴリがあまりにも細かくなってしまうと, カテゴリモードを使用する時, 逆に, 探している情報にたどりつきにくくなる可能性がある. そのため, これらの 8 つのカテゴリは, 地域サイトで 1 ページに書かれる程度の分類カテゴリを調査した結果と, Cyber City Case Bank[1] で用いられている分類カテゴリを参考にして決定した.

4.3 特徴語辞書

それぞれのカテゴリに固有な単語や表現である特徴語は, 特徴語辞書として整理した. この特徴語辞書は, 石川県内の全市町村にあたる 41 地域, 54 サイトにあるページをカテゴリごとに分類し, それらのページのタイトルとアンカ文字列, 強調文字列から, それぞれのカテゴリに固有な単語や表現を選び出すことによって作成した. 各カテゴリの特徴語辞書の内容を論文末 (Appendix A) の表 A.1〜表 A.9に示す.

特徴語パターンの中に含まれる ⟨ 地域名 ⟩, ⟨ 地域タイプ ⟩, ⟨ 首長 ⟩ は, それぞれの地域によって変化するパターンである. 以下に, これらのパターンに該当する文字列を, 石川県を例にとって示す.

⟨ **地域名** ⟩ 地域タイプを含む漢字地域名 (ex. 石川県), 地域タイプを含まない漢字地域名 (ex. 石川), 地域タイプを含むひらがな地域名 (ex. いしかわけん), 地域タイプ含まないひらがな地域名 (ex. いしかわ) が該当.

⟨ **地域タイプ** ⟩ その地域のタイプ (ex. 県) が該当.

⟨ **首長名** ⟩ その地域の首長名 (ex. 石川県知事, 県知事, 知事) が該当.

4.4 分類対象ページ

地域サイト内の全てのページを分類すると, 1 つのカテゴリに多くのページが分類されてしまうことがある. これらの全てをユーザに表示してしまうと, ユーザがその中から知りたい情報を探し出す手間が余計にかかってしまう. 同じカテゴリに分類される多くのページは, その中の幾つかのページから容易に辿れることが多く, そのようなページのみを分類するほうが簡単に知りたい情報を探し出せる. ここでは, 無駄なページを分類しないために, ページのカテゴリ分類処理の対象とするページを限定する.

分類対象ページは, 次の 3 つの条件を満たすものとする.

1. 同一地域サイト内で, トップページから距離 2 以下のページ[1]
2. 親ページと一致しないカテゴリに分類される子ページ
3. 親ページの判定文字列に「新着」「索引」などが含まれないページ

以下では, それぞれについて説明する.

- 同一地域サイト内で, トップページから距離 2 以下のページ
  分類カテゴリで分類されるページは, トップページから距離 2 以下に存在することが多く, この条件は, 処理時間の短縮と, トップページから距離の離れた不適切なページを分類しないために設けている. 例えば, この条件では, 地域サイトの構造が図 4.1 のようになっている場合, 距離 2 の『名所』『宿泊』『イベント』までが分類対象となる.

- 親ページと一致しないカテゴリに分類される子ページ
  親ページがあるカテゴリに分類された場合は, その子ページは親ページが分類されたカテゴリの, さらに詳細な分類でのページである場合が多い. このようなページへは, 既に分類されている親ページから辿っていく方が容易にアクセスできる. 逆に, これらのページも分類し, 表示してしまった場合, 選択するページが増え, アクセスしにくくなる恐れがある.

  そこで, 親ページが分類されたカテゴリでの分類処理は, その子ページでは行なわない. ただし, 分類処理はそれぞれのカテゴリで行なうため, 親ページが分類されなかったカテゴリでは, 子ページでも分類処理を行なう.

[1] ハイパーリンクを 1 回辿ることを距離 1 とする.

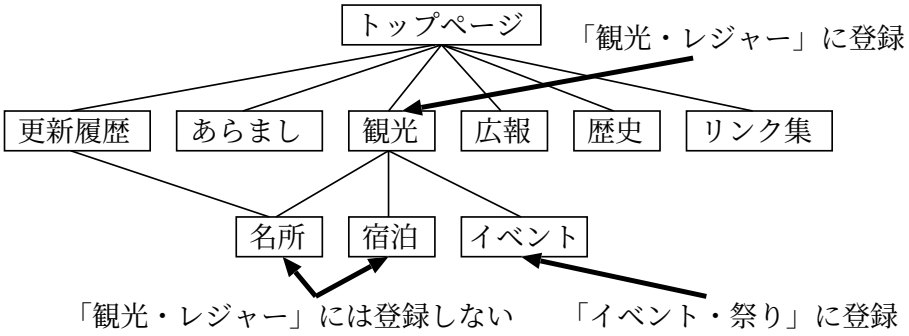

図 4.1: 地域サイトにおけるページ分類

例えば, 図 4.1の地域サイトを考える. カテゴリ「観光・レジャー」に分類されたページ（『観光』）の子ページである『名所』『宿泊』は, 親ページがカテゴリ「観光・レジャー」に分類されなければ, カテゴリ「観光・レジャー」に分類されるが, 親ページがカテゴリ「観光・レジャー」に分類されたため, 分類処理は行なわず, データベースには登録しない. 一方, 親ページのカテゴリと異なるカテゴリでは処理を行なうため, 子ページの『イベント』はカテゴリ「イベント」に分類され, データベースに登録する.

- 親ページの判定文字列に「新着」「索引」などが含まれないページ
  地域サイトによくある「新着情報」や「索引」「更新履歴」といったページからアンカで参照しているページは, 他のページから容易に辿れるものや, 過去の情報を掲載しているページなどが多く, 知りたい情報へのアクセスの妨げになる可能性がある. このため, 「新着」「索引」「更新履歴」「english」「更新状況」「what と new」の 6 語が判定文字列であるアンカ文字列とタイトルに含まれる場合は, その子ページを分類対象としない. ただし, 強調文字列に, それらの語が含まれる場合は, その強調文字列がそのページの内容を指している可能性が高くないためここでの対象としない.

  例えば, 図 4.1の地域サイトでは, ページ『更新履歴』の子ページ『名所』は, 先に分類しているページ『観光』から容易に辿れるページである. このため, 『更新履歴』の子ページは分類対象としない.

4.5 判定文字列

地域情報ページの自動分類では, ページをどのカテゴリに分類するかを判定する手掛りとして, 次の 3 種類の文字列を用いている.

1. アンカ文字列 · · · アンカ開始タグ (⟨a⟩) とアンカ閉じタグ (⟨/a⟩) で囲まれた文字列. この文字列が対象としているページは, アンカ文字列が掲載されているページではなく, アンカが指す URL に対応するページである. 人間はアンカ文字列を見て, そのリンク先のページを見るかどうか判断することが多い. このため, アンカ文字列には, リンク先のページの内容を正確に, 短くまとめた文字列が使われる可能性が非常に高いと言える. ただし, アンカ文字列の代わりに図を入れることも可能なので, 全てのアンカからアンカ文字列が抽出できるわけではない.

2. ページタイトル · · · タイトル開始タグ (⟨title⟩) とタイトル閉じタグ (⟨/title⟩) で囲まれた文字列. 対象としているページは, ページタイトルが掲載されているページである. その名の通り, そのページの内容を短く要約した文字列であることが多い. しかし, あまり人目につきやすいとは言えず, ページから情報を得る際にもそれほど重要ではないため, タイトルのないページがあったり, サイトの全てが同じタイトルを持っていたりすることもしばしばある.

3. 強調文字列 · · · ページ内で見出しなどに使われている文字列. 他の 2 つの文字列は, HTML のタグ情報から簡単に抽出できるのに対し, この文字列は抽出のために複雑な処理を行なう. この文字列は, 人目につきやすい文字列であるが, それらがそのページ全体の内容を表現している可能性は, 他の 2 つの文字列より低い. ここでは, 他の 2 つの文字列の補助的なものとして使う.

4.5.1 強調文字列の抽出

アンカ文字列とタイトルは, HTML のタグ情報から容易に取り出すことができる. しかし, 強調文字列は抽出するために, 幾つかの処理をする必要がある.

ここでは, 文字列を強調するために良く使われるタグ (以後, 文字列強調タグ) として, ⟨font⟩, ⟨h1⟩, ⟨h2⟩,⟨h3⟩, ⟨h4⟩, ⟨h5⟩,⟨h6⟩, ⟨b⟩, ⟨em⟩,⟨strong⟩, ⟨big⟩ を設定する. また, リストの内容も文字列を強調できるため, ⟨ol⟩,⟨ul⟩, ⟨dl⟩ も上記のタグと同様に, 文字列強調タグとして設定する.

これらの文字列強調タグによって強調されていると思われる文字列を, 全て, 判定文字列としたのでは膨大な処理時間がかかってしまう. このため, ここでは強調されていると

思われる文字列の文字列強調タグによってスコアを付け, 1 ページにつき, スコアの高い 3 つの文字列を強調文字列として抽出する.

スコア付けの際に問題となるのは, 文字列強調タグが入れ子になっているため, 文字列にどのような文字列強調タグがついているのかわかりにくいということである.

単純に, ある文字列強調タグのスコープ (有効範囲) を文字列として抽出すると, 抽出した文字列の中に短い強調文字列があったり, 外側の文字列強調タグのスコープであるにも関わらず, その文字列強調タグが抽出できないといったことがおこる. このため, 文字列強調タグのスコープの中で最も短いものを基準として, 外側の文字列強調タグのスコープである場合は, 区切った文字列の直前と直後の両方に, 外側の文字列強調閉じタグと開始タグを追加するという前処理を行なう.

また, ここでは見出しとして強調されている文字列を抽出したいため, 見出しである可能性の高い, 50Bytes 以下の文字列を抽出する.

この処理は, 有賀 [6] の方法を参考にした, 次の方法で行なう.

1. 入れ子になっている強調文字列を取り出しやすくするため前処理を行なう.

2. 設定した文字列強調タグの, 開始タグと閉じタグの間にある文字列を, 文字列強調タグと共に抽出する. ここでは, 1 つの文字列に対して, 複数の文字列強調タグを抽出する場合もある.

3. 抽出した強調文字列の中から, アンカ文字列を消去する. アンカ文字列は, アンカのリンク先のページに関係する文字列である可能性が非常に高く, アンカ文字列が存在するページの内容を表わしているものではないため, ここで消去する.

4. 抽出した強調文字列の中から, 50Bytes 以下の文字列のみを残す.

5. 残った文字列のスコアを計算する. スコアの計算方法は, 以下の通りである.

   ⟨**font**⟩ size 属性が設定されている場合は, 標準サイズを 3 として, その差分をスコアに加算する. 例えば, size の値が 5 の場合は, $5 - 3$ の 2 点を加算する. また, color 属性が設定されている場合は, 1 点をスコアに加算する.

   ⟨**h1**⟩**〜**⟨**h6**⟩ ⟨h5⟩ を標準として, 数字の差分をスコアに加算する. ⟨font⟩ と違って, 数が少ない方が良く目立つため, $5 - [h$ **につづく数字**$]$ という計算でスコアを決定する. 例えば, ⟨h2⟩ は, $5 - 2$ で, 3 点が加算される.

   ⟨**b**⟩, ⟨**em**⟩, ⟨**strong**⟩, ⟨**big**⟩, ⟨**ol**⟩, ⟨**ul**⟩, ⟨**dl**⟩ それぞれ, 1 点ずつ加算する.

また, 下記の特殊文字で始まる文字列は, 良く目立ち, 見出しのはじめの文字として使われやすいため, これらの文字で始まる文字列には, スコアに 1 点を加算する.

＊@§☆★○●◎◇◆□■△▲▽▼【―・〔＜［

6. スコアが 0 以下の強調文字列を消去する.

7. 1 ページに 4 つ以上の強調文字列が残っている場合は, スコアが高い 3 つの文字列だけを強調文字列として抽出する. また, スコアが同じ文字列がある場合は, ページのはじめの方に書かれているものを優先的に選択する.

4.5.2 分類処理の優先順位

アンカ文字列, タイトル, 強調文字列のいずれかに, カテゴリの特徴語パターンが含まれているページは無条件に, そのカテゴリに分類されるわけではない. 処理時間短縮と分類するには不適合なページをできるだけ分類しないようにするため, これら 3 種類の文字列に, 次のような優先順位を持たせて処理を行なう.

アンカ文字列 〉ページタイトル 〉強調文字列

ページの分類処理は, 優先順位の高い判定文字列から分類の手掛りとして用いられる. 例えば, あるカテゴリの分類処理で, 優先順位が上位の判定文字列を手掛りとしてページが分類された場合, 下位の判定文字列を手掛りとして同じカテゴリの分類処理は行なわない. ここで分類の手掛りとして使われた判定文字列は, カテゴリ分類テーブルの文字列フィールドに登録される. そのため, 同一カテゴリの分類処理での判定文字列の優先順位は, 文字列フィールドに登録される優先順位でもある.

また, 同一カテゴリ以外の分類処理にも, この優先順位に従って制限を加える. 例えば, あるページがアンカ文字列を手掛りとして, カテゴリ「観光・レジャー」に分類されたとする. この時, 下位の判定文字列であるページタイトルと強調文字列は, 同一カテゴリの「観光・レジャー」の分類処理を行なわない. しかし, 「統計」など同一カテゴリ以外の分類処理はアンカ文字列とページタイトルで行ない, 強調文字列は行なわない. 前述のように強調文字列は抽出の際, 多くの処理が必要で時間がかかることと, 強調文字列の信頼性が低いためこのような方法をとる.

上位の判定文字列を手掛りとしてページが分類されたときの, それぞれの判定文字列の分類処理を表 4.1にまとめる.

もし, 特徴語がアンカ文字列, ページタイトル, 強調文字列に 1 つも含まれない場合は, 分類不能ページとして分類しない.

表 4.1: 上位の判定文字列でページが分類されたときの分類処理

判定文字列	分類されたカテゴリの分類処理	分類されていないカテゴリの分類処理
アンカ文字列	—	行なう
ページタイトル	行なわない	行なう
強調文字列	行なわない	行なわない

4.6 自動分類アルゴリズム

以上の方法に基づき, 1 つの地域サイトにあるページの, カテゴリ分類の処理は, 次のようにして行なう.

1. その地域サイトのトップページから, 分類対象となるページを, 幅優先探索で順に取り出す.

2. 取り出したページに対して, 判定文字列の優先順位の高いものから, 次の処理を行なう.

   (a) 対象となる文字列を抽出する.

   (b) 抽出した文字列の地域名, 地域タイプ, 首長を, それぞれ ⟨ 地域名 ⟩, ⟨ 地域タイプ ⟩, ⟨ 首長 ⟩ に置き換える.

   (c) 分類対象となるカテゴリの, 特徴語辞書に登録されている特徴語パターンが, その文字列に含まれるかを 1 パターンずつ調べる. 特徴語パターンが, 文字列の中に含まれていたときは, そのページの URL と判定文字列, 特徴語, 分類したページがある地域サイトが対象としている地域の地域文字列を 1 レコードとしてカテゴリ分類テーブルに登録する. また, そのカテゴリの, 残りの特徴語パターンの調査は行なわない. この処理は, 対象となる全てのカテゴリに対して, 同様の処理を行なう.

   (d) 分類対象となる全てのカテゴリに対してこれらの処理が終了し, このページが, いずれかのカテゴリに分類された場合, 優先順位の低い文字列に対して, そのカテゴリでの処理を行なわない. また, 分類されなかったカテゴリの分類処理はページタイトルに対しては行なう. いずれのカテゴリにも分類されなかった場合は, 優先順位が次の文字列に対して, 処理 (a) から同様の処理を行なう. 現在の処理で対象としている文字列が, 最も優先順位の低い文字列であった場合

表 4.2: 自動分類の実験結果 (Closed テスト)

分類カテゴリ	分類すべきページ数	実際分類したページ数	分類正解ページ数	再現率	適合率
一般	65	69	54	83.1%	78.3%
計画・産業	50	59	49	98.0%	83.1%
イベント・祭り	128	117	110	85.9%	94.0%
文化・歴史・教育	163	162	146	89.6%	90.1%
観光・レジャー	286	284	248	86.7%	87.3%
統計	12	10	9	75.0%	90.0%
住民向け	103	107	95	92.2%	88.8%
リンク	22	21	18	81.8%	87.9%
合計	829	829	729	87.9%	87.9%

は, そのページを, 分類不能ページとし, そのページはこれ以上の処理を行なわない.

4.7 実験と検討

4.7.1 実験

上記の方法に基づき, ページを自動分類する実験を行なった. 実験は, 特徴語辞書の作成で使用したサイト（石川県内の 41 地域の 54 サイト）を対象とした Closed テストと, 全国 27 地域の 33 サイトを対象とした Open テストの両方を行なった. Closed テストの結果を表 4.2, Open テストの結果を表 4.3に示す.

Closed テストでは, 人が分類するべきと判断したページ数が 829 ページ, 実際に自動分類したページ数も 829 ページであった. そのうち, 分類正解ページ数は 729 ページで, 再現率と適合率の両方が 87.9%となった. Open テストでは, 分類するべきと判断したページ数が 962 ページ, 自動分類したページ数が 902 ページであった. 分類正解ページ数は 687 ページで, 再現率が 71.4%, 適合率が 76.2%となった.

また, 判定文字列ごとの適合率を求めた. Closed テストの判定文字列ごとの適合率を表 4.4, Open テストの判定文字列ごとの適合率を表 4.5に示す.

アンカ文字列を手掛りとして分類したページの適合率は, Closed テストで 91.0%, Open テストで 88.4%である. また, ページタイトルを手掛りとして分類したページの適合率は,

表 4.3: 自動分類の実験結果 (Open テスト)

分類カテゴリ	分類すべきページ数	実際分類したページ数	分類正解ページ数	再現率	適合率
一般	55	55	47	85.5%	85.5%
計画・産業	60	72	51	85.0%	70.8%
イベント・祭り	100	86	75	75.0%	87.2%
文化・歴史・教育	149	128	87	58.4%	68.0%
観光・レジャー	263	248	187	71.1%	75.4%
統計	30	19	19	63.3%	100.0%
住民向け	287	274	204	71.1%	74.5%
リンク	18	20	17	94.4%	85.0%
合計	962	902	687	71.4%	76.2%

表 4.4: 判定文字列の適合率 (Closed テスト)

判定文字列	実際分類したページ数	分類正解ページ数	適合率
アンカ文字列	634	577	91.0%
ページタイトル	152	123	80.9%
強調文字列	43	29	67.4%

表 4.5: 判定文字列の適合率 (Open テスト)

判定文字列	実際分類したページ数	分類正解ページ数	適合率
アンカ文字列	611	540	88.4%
ページタイトル	85	68	80.0%
強調文字列	206	79	38.3%

Closed テストで 80.9%, Open テストで 80.0%である. この 2 つの文字列は, Closed テストと Open テストの両方で安定した適合率である. 強調文字列を手掛りとして分類したページの適合率は, Closed テストで 67.4%, Open テストで 38.3%で, 他の 2 つの文字列より精度が劣っている.

4.7.2 検討

表 4.2と表 4.3に示す分類精度の実験では, Open テストの再現率が 71.4%とやや低かったものの, 比較的単純な分類方法にしては良い精度が得られた.

間違ってページを分類した原因は, まず, 強調文字列にページの内容と関係のない文字列が混ざっていたということが挙げられる. 例えば, 「観光」に分類するべきページに “○○市産業観光振興課” という文字列が強調されて掲載されていて, そのページが他の判定文字列で分類されなければ, 間違ったカテゴリである「計画・産業」に分類されてしまう.

他にも, ある地域を対象とする地域サイトと同じサイトに, 他地域の地域サイトがあり, 対象とする地域に関する記述が 1 ページしか無い場合は, 他の地域のページを分類してしまうことがある. これは, 他の地域の地域サイトまでの距離が 2 より離れていないために起こる.

分類すべきページを分類しなかった原因は, 特徴語の不備がほとんどであるが, 中には 3 つの判定文字列では分類できないページもあった.

また, 判定文字列が固有名詞のみのページがあり, これらのほとんどは分類されていない.

表 4.4と表 4.5に示す, 分類の手掛りとした判定文字列ごとの適合率の調査実験は, 既に優先順位を決定して分類処理を行なった結果であるため, 確実な証拠とはいえないが, ページの内容を知る際に, アンカ文字列とページタイトルは非常に有用な情報になることが推測できる. 特に, アンカ文字列を判定文字列として分類して, それが正しかったページは, 全正解ページの約 80%を占めた.

一方, アンカ文字列もページタイトルも存在しないというページはしばしば見られる. この場合, ページ内の強調文字列は分類の有効な手掛りとなる. 実際, 適合率はそれほどよくなかったものの, Open テストの分類正解ページの 10%以上が強調文字列を手掛りとして分類されている.

文献 [12] では, 判定文字列をアンカ文字列とページタイトルのみとした, 地域情報ページの自動分類の実験について述べた. 今回の分類実験と, 前回の分類実験における分類方式の違いは以下の 3 つである.

1. ⟨ 地域名 ⟩ など抽象的パターンの導入

表 4.6: 前回の分類実験結果と今回の分類実験結果の比較

	Closed テスト		Open テスト	
	再現率	適合率	再現率	適合率
前回	77.9%	85.8%	67.1%	80.3%
今回	87.9%	87.9%	71.4%	76.2%
●	84.4%	89.1%	63.2%	87.4%

2. 強調文字列を判定文字列へ採用

3. 特徴語辞書の増補と修正

この 3 つのうち, 1 の ⟨ 地域名 ⟩ など抽象的パターンの導入と, 3 の特徴語辞書の登録数の修正は, 適合率を改善することを目的として行なった. また, 2 の強調文字列を判定文字列へ採用と, 3 の特徴語辞書の増補は再現率を改善することを目的として行なった.

前回の実験結果と今回の実験結果の比較を, 表 4.6に示す. Closed テストでは再現率と適合率の両方が向上している. これは, 新たに導入した上記の 3 点により改善できたと考えられる.

一方, Open テストでは, 再現率は向上したが, 適合率が悪くなっている. 今回の Open テストの実験結果を用いて, 強調文字列を判定文字列としないで分類した結果を計算すると, 表 4.6の●に示す結果となる. これにより, 今回の実験での適合率の低下は強調文字列を判定文字列へ採用が大きく原因していると考えられる. しかし, 強調文字列を判定文字列へ採用しなかった場合の再現率が 63.2%のままでは, このシステムが実用的とは言えないことを表わしている. そのため, 強調文字列の適合率を向上させる何らかの方法を考える必要がある.

第 5 章

特定情報サイトからの情報抽出と情報生成

本章では, 特定情報サイトのページにある表から地域情報を抽出する**情報抽出**と, 抽出した数値情報を組み合わせて新たな情報を生成する, **情報生成**について説明する. 情報抽出・生成処理によって得た地域情報は, コンテンツデータベースの特定情報テーブルに登録する.

特定情報サイトには, 人口や面積などの有用な地域情報が存在する. これらの地域情報は表形式で掲載されている場合が多く, ほとんどの地域の情報が提供されている. このため, このようなページからの情報抽出と情報生成処理を行なうことで, 地域サイトが収集されなかった地域でも, 何らかの地域情報を提供できる.

5.1 情報抽出

情報抽出では, 表 5.1に示す特定情報サイトにある表形式で提供されている地域情報を抽出する. 抽出した情報はコンテンツデータベースの特定情報テーブルに登録する.

この処理では, 情報源となる特定情報サイトのトップページの URL はあらかじめ登録

表 5.1: 情報抽出を行なう特定情報サイト

項目	特定情報サイト	URL
人口・世帯数	平成７年国勢調査	http://www.stat.go.jp/0513.htm
面積	都道府県別面積	http://www.gsi-mc.go.jp/MAP/MENCHO/ichiran.htm
役所情報	地方公共団体住所一覧	http://www.lasdec.nippon-net.ne.jp/jyuusyo/jyu_top.htm

図 5.1: 特定情報サイトのトップページ (目次ページ)

しておく. その特定情報サイトのトップページには, 次の 2 種類がある.

- 目次ページ：都道府県名をアンカ文字列としたアンカが存在する. 多くの場合, そのリンク先ページに, 都道府県の各地域の情報を記載した表が存在する. 図 5.1で示す地方公共団体コード住所一覧サイトのトップページがこれに該当する.

- 都道府県に対する情報ページ：都道府県レベルの地域に対する情報を表の形で提示している. その表の中の都道府県名, もしくは道東, 道北, 道央, 道南などの地方名にハイパーリンクが埋め込まれており, そのリンク先ページは, 都道府県の各地域の情報を記載した表が存在する. 第 1 章で示した都道府県別面積サイトのトップページ (図 1.2) がこれに該当する.

また, 特定情報の表は, トップページから距離 2 以内のページに掲載されている. そこで, 地域名か, 道東, 道北, 道央, 道南といった北海道の地方名を含むアンカ文字列のアンカでリンクされ, トップページから距離 2 までのページにある, すべての表を解析し, 地域情報と判定されたセルの情報を, すべて, 特定情報テーブルに登録する.

具体的なアルゴリズムを次に示す.

1. 与えられた URL から, その URL に対応するページを取得する.

2. そのページに表が存在した場合は, 山本ら [9] の方法に従ってその表を解析し, それぞれの地域に対するレコードを抽出する.

3. 抽出したレコードの全ての値とそのフィールド名を, 値は属性値, フィールド名は属性として, 対象とする地域の地域コードと一緒に特定情報テーブルに登録する. また, 表があるページの URL も登録する.

4. そのページに地域名か, 地方名を含むアンカ文字列のアンカが存在した場合は, そのURL を抽出して, 1 から 4 を繰り返す. この処理は, 対象ページがトップページから距離 2 以内の場合に行なう.

表に掲載されている属性値は全て特定情報テーブルに登録し, 表示モジュールで必要なものだけを検索するようにしている. そのため, 表 2.7に示すように, 地域情報として表示をしない「市区町村番号」と「団体コード」もテーブルに登録している.

5.2 情報生成

情報生成では, 抽出した地域情報の数値データから, 新たな情報を生成する. この処理で現在生成している情報は以下の 2 種類である.

- 人口密度 = 人口/面積
  同じ地域の, 国勢調査サイトから得られた人口と, 都道府県別面積サイトから得られた面積を特定情報テーブルから取り出し, 上記の計算を行ない値を出す.

- 人口, 世帯数, 面積, 人口密度の全国での順位と, 同一都道府県内での順位
  全国での順位は, 全地域を都道府県, 市町村, 区の 3 つのレベルに分けて, その地域の該当するレベルにある, 全国の地域の中での順位を出す. 同様に, 同一都道府県内での順位は, 市町村, 区の 2 つのレベルに分けて, 該当するレベルにある, 同一都道府県内の地域の中での順位を出す. ここでは, SQL の order(並べ替え) 関数を用いてレコードを整列させ, 並んだ順番に順位をつける.

生成した情報はコンテンツデータベースの特定情報テーブルに登録する.

具体的なアルゴリズムを次に示す.

- 人口密度
  すべての地域に対して, 次の処理を行なう.

1. 対象地域の国勢調査サイトから得られた人口と, 都道府県別面積サイトから得られた面積の値があるか調べる. なければ, 人口密度は出さない.
2. 人口/面積 を計算し, その値を属性値, 属性を「人口密度」として特定情報テーブルに登録する.

- 順位
  人口, 世帯数, 面積, 人口密度のそれぞれにおいて, 全国と同一都道府県の両方で, 次の処理を行なう.

  1. 地域を都道府県 (全国の処理のみ), 市町村, 区の 3 つのグループに分け, それぞれに該当するすべてのレコードを, 属性値で並べ替える.
  2. 該当するレコードの総数を求める.
  3. 順位フィールドに, グループ内で値の大きなものから順に順位を登録する. 同時に, 地域総数にレコードの総数も登録する.

第 6 章

検討および関連研究

6.1 検討

本システムで生成するディレクトリは, 以下の特長を持つ.

- 日本全国の 47 都道府県と 3380 の市区町村の合計 3427 自治体のほぼ全てで, 何らかの地域情報を提供する.
- 地域サイト特定情報サイトの 2 種類の情報源を利用するため, 掲載している情報が豊富である.
- 地域サイトは自動的に収集, 分類され, 更新が容易に行なえるため, 最新の情報が提供できる可能性が高い.
- 地域モードとカテゴリモードの 2 種類の表示方法を提供しているため, 用途に合った使い方を可能とする.
- 情報提供の書式が統一されているため, 比較的容易に必要な情報にアクセスできる.

このような特長から, 本ディレクトリは旅行に行く際などに有用な情報源となり得る. また, 最近, 小学校でインターネットを用いた教育が導入されつつあるが, そのための資料作りはかなり手間がかかることである. この際にも, 本ディレクトリは教材として有用であると考えられる.

また, 本システムの中心技術は**地域サイトの自動収集**と**地域情報ページの自動分類**である.

地域サイトの自動収集では, 提案手法によって, 全体の 80%以上の地域で情報源となる地域サイトを収集することができた. このように多数のサイトを収集できたのは, 次のような理由によると考えられる.

1. 地域サイトであることが確実な URL が入手可能である（簡単な規則によって生成できる）. この URL を用いて, リンク集を特定することができる.

2. 地域サイトへのリンクを持つ, 地域情報関連のリンク集が, 多数, 存在している. 地域サイトへのハイパーリンクのアンカ文字列には, ほとんどの場合「地域名」が用いられるため, これらのリンク集から地域サイトの URL を安定して抽出できる.

これらの条件は, 地域情報以外の情報収集においては, 一般に成り立たない. このため, 提案手法をそのままの形で他の領域に適用することはできない. しかし, 情報収集にリンク集を利用するという方法は, かなり強力な方法であり, Clever サーチ [7] や佐藤らのリンク集の自動生成 [10], 山本らの人物情報の自動収集 [8][9] でも有効に働くことが報告されている. 上記の (1) は, 既知の URL がある程度入手可能であれば, それを種としてブートストラップ的にリンク集を見つけることができることを示している.

一方, 地域情報ページの自動分類では, 地域情報ページのタイトルやアンカ文字列, ページ内の強調文字列に見られる定型的な表現に着目し, 比較的, 簡便な方法でページの分類を実現した. ここでは, Open テストの再現率がやや低かったもののまずまずの結果が得られた. この理由として, 次の 2 つが考えられる.

1. 地域情報のみを対象としているため, カテゴリ特有の表現や単語の数が限られている. また, それらが曖昧性を持つことが少ない.

2. ページの内容がアンカ文字列とタイトル, 強調文字列だけで理解できることが多い. 特に, アンカ文字列は, そのページの内容を, よく表していると言える.

1 つ目の理由は地域情報ページ以外で有効であるかはわからないが, ある程度分野が限定されているなら, 有効であると予想できる.

2 つ目のアンカ文字列がページの内容を知るための有効な情報源となっていることは, 佐藤らのリンク集の自動生成 [10], 山本らの人物情報の自動収集 [8][9] でも報告されている.

6.2 関連研究

本研究に最も関連した研究は, リンク集の自動生成に関する研究である.

Clever プロジェクト [7] は, サーチエンジンの高度化を目的としたプロジェクトで, ある入力（例えば, “cheese”）に対して, それに関する少数, 信頼できるページを得る方法を実現している. これらのページは, いわゆるリンク集に相当する hubs と, 多くの hubs からリンクされているページ (authorities) から成る. HITS アルゴリズムは, これら 2 種類のページを, その間の依存関係を利用した繰り返し計算によって見つける方法を与えている. 我々の情報収集の方法はこの方法と似ているが, 他の方法で authorities を見つけることができるため, より簡単な方法でリンク集を見つけることができる.

佐藤ら [10] は, カテゴリ名を入力として, そのカテゴリに対するリンク集を自動生成する方法を提案している. この方法は, まず, カテゴリ名（例えば, 「水族館」）からそのカテゴリに属するインスタンス名（例えば, 「おたる水族館」）を収集し, 次に, 見つかったインスタンスに対する情報を収集することによってリンク集を作成する方法をとっている. 地域情報の場合は, あらかじめすべての地域名（自治体名）がわかっているため, このような方法をとる必要がない.

本研究以外に, WWW に対する自動処理によって実現されている地域情報提供システムには, モーバイルインフォサーチ [11] がある. この研究では, 特に位置情報に着目してシステムを構成し, その情報源として, 地図やイエローページなど検索機能を有するデータベースタイプのサイトと, 位置情報の記述している一般的なページの 2 種類を利用している. このシステムは, 例えば, 住所（の一部）から, その近くにあるお店のページなどが簡単に検索できる. これに対して本研究は, 自治体で分割した地域を対象として, 地域毎に情報を整理して提示することを目的としている.

第 7 章

結論

本論文では, 地域情報を対象としたウェブディレクトリとそれを自動生成するシステムを提案した.

本システムは, コンテンツデータベース, 表示モジュール, 地域情報編集モジュールの 3 つの要素から構成される. コンテンツデータベースは, 本地域情報ディレクトリで提供する全ての情報を格納したデータベースである. 表示モジュールは, ユーザの要求に合ったディレクトリを, コンテンツデータベースの内容から生成し提供する. 本ディレクトリは, 地域別に情報を表示する地域モードと, 複数の地域の特定カテゴリに対するページを表示するカテゴリモードの 2 つの表示方法を提供する.

本システムの主要部となる地域情報編集モジュールでは, 地域サイトの自動収集, 地域情報ページの自動分類, 特定情報サイトからの情報抽出と生成を行なう.

地域サイトの自動収集では, 地域サイトの URL の典型的なパターンを用いて地域サイトのリンク集を発見し, そのリンク集から地域サイトのトップページの URL を収集した. この方法により全体の 83.2%の 2852 地域に対して, 地域サイトを 1 つ以上発見することができ, 全部で 4012 サイトもの地域サイトを収集することができた. これは, 非常に多くの地域サイトを掲載しているリンク集を上回る収集数である. このため, この方法での地域サイトの自動収集は, 実用レベルに達していると言える.

また, 地域情報ページの自動分類では, アンカ文字列, ページのタイトル, ページ内の強調文字列に現れる特徴的な表現を利用して, 「一般」「計画・産業」「イベント・祭り」「文化・歴史・教育」「観光・レジャー」「統計」「住民向け」「リンク」の 8 つのカテゴリにページを分類した. この方法により, Closed テストでは, 再現率, 適合率とも 87.9%, Open テストでは, 再現率 71.4%, 適合率 76.2%と, 比較的良い結果が得られた. この処理により, カテゴリモードの表示が可能となる. 地域情報ページの自動分類は, 結果はそこそこ良かっ

たものの, 幾つか改善する余地がある.

特定情報サイトからの情報抽出と生成では, 特定情報サイトのページにある表から地域情報を抽出し, それらの情報を組み合わせて新たな情報を生成する. これにより, 地域サイトが開設されていない地域にも何らかの情報を提供することが可能となった.

本ディレクトリが地域情報へのアクセス支援として使われることにより, WWW 上の地域情報が有効に利用できる. それにより, WWW から地域情報を提供することが有効であると認識され, WWW 上の地域情報が, ますます充実していくことを期待する.

謝辞

本研究を通して多くの御教示を頂きました佐藤理史助教授に心から感謝致します.

また, 日頃より研究に関するアドバイスを下さった佐藤研究室の皆様に心より感謝致します.

そして, 北陸先端大での学生生活を様々な面から支えてくださった両親, および, 友人達に深く感謝致します.

参考文献

[1] NRI サイバー都市ケースバンク, サイバー社会基盤研究推進センター: Cyber City Case Bank, http://www.ccci.or.jp/city-cb/.

[2] 日本規格協会: 都道府県コード (JIS X0401), JIS ハンドブック 情報処理 用語・符合・データコード編 - 1999, pp764, 1999.

[3] 日本規格協会: 市区町村コード (JIS X0402), JIS ハンドブック 情報処理 用語・符合・データコード編 - 1999, pp765–840, 1999.

[4] 日本ネットワークインフォメーションセンター: ドメイン名登録等に関する技術細則, http://www.nic.ad.jp/jpnic/domain/saisoku-1.html.

[5] 石田和生, 谷川哲司, 宮下敏昭: WWW におけるダングリングリンクの自動メンテナンス. 第 59 回情報処理学会全国大会, Vol.3, pp85–86, 1999.9.

[6] 有賀忠徳: カテゴリに基づく製品情報の組織化. 北陸先端科学技術大学院大学修士論文, 2000.

[7] Members of the Clever Project: Hypersearching the Web. *Scientific American*, Vol.280, No.6, pp54–60, 1999.

[8] 山本あゆみ, 佐藤理史: WWW 上の職業別人名リストを利用した人名の収集. 情報処理学会第 59 回全国大会, Vol.3, pp119–120, 1999.

[9] 山本あゆみ, 佐藤理史: ワールドワイドウェブからの人物情報の自動収集. 第 119 回情報処理学会「知能と複雑系」研究会 (ICS-119), pp173–180 2000.

[10] Satoshi Sato and Madoka Sato: Toward Automatic Generation of Web Directories. *Proc. of International Symposium on Digital Libraries 1999 (ISDL'99)*, pp127–134, 1999.

[11] 三浦信幸, 高橋克巳, 横路誠司, 島健一: 位置指向の情報統合 ～モーバイルインフォサーチ 2 実験～. 第 57 回情報処理学会国大会, Vol.3, pp637–638, 1998.

[12] 大槻洋輔, 佐藤理史: ワールドワイドウェブを知識源とした地域情報の自動編集. 第 119 回情報処理学会「知能と複雑系」研究会 (ICS-119), pp165–172, 2000.

Appendix A

特徴語辞書

地域情報ページを分類する際に使う特徴語を特徴語辞書としてまとめた. その内容を表A.1〜表 A.9に示す.

表 A.1: 一般の特徴語辞書 (正規表現)

1. ⟨ 首長 ⟩(から)?の?(一言 | ひとこと)
2. ⟨(首長 | 地域名)⟩(から)?の?(メッセージ | めっせーじ)
3. ⟨ 首長 ⟩ の?(挨拶 | あいさつ)
4. ⟨ 地域タイプ ⟩(章 | しょう)
5. ⟨(地域タイプ | 地域名)⟩ の?シンボル
6. シンボルマーク
7. ⟨(地域タイプ | 地域名)⟩ の?木
8. ⟨(地域タイプ | 地域名)⟩ の?(花 | はな)
9. ⟨(地域タイプ | 地域名)⟩ の?(鳥 | とり)
10. 名木
11. 名水
12. 天然記念物
13. ⟨(地域タイプ | 地域名)⟩ 名の由来
14. (姉妹 | しまい)(都市 | とし)
15. (友好 | ゆうこう)(都市 | とし)
16. 国際交流
17. 地勢
18. 地理
19. アクセス
20. 所要 (時間 | じかん)
21. ⟨(地域タイプ | 地域名)⟩ までの交通
22. 気候
23. (^|⟨(地域タイプ | 地域名)⟩ の?) あらまし
24. ⟨(地域タイプ | 地域名)⟩ の?(概要 | がいよう)
25. outline
26. (^|⟨(地域タイプ | 地域名)⟩ の?)(プロフィール | プロフィル |profile| ぷろふぃーる)
27. (^|⟨(地域タイプ | 地域名)⟩ の?)(沿革 | えんかく)
28. ⟨(地域タイプ | 地域名)⟩ の?ガイダンス
29. ⟨(地域タイプ | 地域名)⟩ について
30. ⟨(地域タイプ | 地域名)⟩ の?ご?(紹介 | しょうかい)
31. ⟨(地域タイプ | 地域名)⟩ の?情報
32. ⟨(地域タイプ | 地域名)⟩ の?(歩み | あゆみ)

表 A.2: 計画・産業の特徴語辞書 (正規表現)

1. (プロジェクト | ぷろじぇくと)
2. (プラン | ぷらん)
3. (構想 | こうそう)
4. (計画 | けいかく)
5. 策定
6. (工場 | こうじょう) 立地
7. 分譲
8. (サイエンス | さいえんす)(パーク | ぱーく)
9. (テクノ | てくの)(パーク | ぱーく)
10. (農業 | のうぎょう)
11. 栽培法
12. (工業 | こうぎょう)
13. (都市 | とし)(作り | づくり)
14. (町 | 街 | まち)(作り | づくり)
15. (生活 | せいかつ)(環境 | かんきょう)
16. (商店 | しょうてん)(街 | がい)
17. (研究 | けんきゅう)
18. (整備 | せいび)
19. 建設
20. 着工
21. (事業 | じぎょう)
22. (産業 | さんぎょう)

表 A.3: イベント・祭の特徴語辞書 (正規表現)

1. (大会 | たいかい)($|\s)
2. (鑑賞 | かんしょう)
3. 観賞
4. 展 ($|\s)
5. 一般公開
6. (歳時記 | さいじき)
7. (イベント |event| いべんと)
8. (行事 | ぎょうじ)
9. (事業 | じぎょう)(予定 | よてい)
10. 成人式
11. (祭 | 祭り | まつり)
12. スケジュール
13. スポレク
14. (御輿 | みこし)
15. トライアル
16. シンポジウム
17. (フェア | フェスタ | フェスティバル)($|\s)
18. コンペ
19. (開催 | かいさい)
20. (催し | もよおし)

表 A.4: 文化・歴史・教育の特徴語辞書 (正規表現)

1. パソコン教室
2. (^|\s)(能 | のう)($|\s)
3. (狂言 | きょうげん)
4. (獅子 | しし)(舞 | まい)
5. (浄瑠璃 | じょうるり)
6. (芸能 | げいのう)
7. (演劇 | えんげき)
8. 絵画
9. 文庫
10. 発刊
11. 作品
12. (俳句 | はいく)
13. (論文 | ろんぶん)
14. (昔 | むかし)(昔 | むかし)
15. ミステリー
16. (著名人 | ちょめいじん)
17. (偉人 | いじん)
18. (昔 | むかし)(話 | ばなし | はなし)
19. (方言 | ほうげん)
20. (民話 | みんわ)
21. (伝説 | でんせつ)
22. (伝承 | でんしょう)
23. (史跡 | しせき)
24. (遺跡 | いせき)
25. (古墳 | こふん)
26. (文学 | ぶんがく)
27. (工房 | こうぼう)
28. (博物 | はくぶつ)(館 | かん)
29. ミュージアム
30. 風土
31. 芸術
32. 美術
33. (歴史 | れきし)
34. ⟨(地域タイプ | 地域名)⟩ の?(歩み | あゆみ)
35. (文化 | ぶんか)
36. (伝統 | でんとう)
37. (教育 | きょういく)

表 A.5: 統計の特徴語辞書 (正規表現)

1. (人口 | じんこう)
2. (面積 | めんせき)
3. 指標
4. 指数
5. (数字 | すうじ) で (見 | み) る
6. (数 | かず) の (推移 | すいい)
7. (統計 | とうけい |toukei)
8. (データ | でーた)

表 A.6: リンクの特徴語辞書 (正規表現)

1. (リンク | りんく |link)

表 A.7: 観光・レジャーの特徴語辞書：その 1(正規表現)

1. (アクセス | あくせす)	36. 霊場
2. (マップ | まっぷ |map) で (紹介 | しょうかい)	37. (涅槃 | ねはん)(像 | ぞう)
3. (イラスト | いらすと)(マップ | まっぷ |map)	38. 地図
4. (エリア | えりあ)(マップ | まっぷ |map)	39. (史跡 | しせき)
5. (タウン | たうん)(マップ | まっぷ |map)	40. 文化財
6. ⟨ 地域名 ⟩(マップ | まっぷ |map)	41. 千枚田
7. 所要時間	42. 波の花
8. 地図から	43. の関 ($|\s)
9. ⟨(地域タイプ | 地域名)⟩ までの交通	44. (美術 | びじゅつ)(館 | かん)
10. (交通 | こうつう)(案内 | あんない)	45. (博物 | はくぶつ)(館 | かん)
11. (スポット | すぽっと)	46. (ミュージアム | みゅーじあむ |museum)
12. (見 | み)(どころ | 所)	47. (水族 | すいぞく)(館 | かん)
13. 見処	48. (記念 | きねん)(館 | かん)
14. (町並 | まちなみ)	49. の (館 | やかた)
15. playland	50. 昆虫館
16. (宿泊 | しゅくはく)	51. 天体観察センタ
17. お (宿 | やど)	52. (動物 | どうぶつ)(園 | えん)
18. 泊ま	53. (工房 | こうぼう)
19. 1 泊	54. (入園 | にゅうえん)(料 | りょう)
20. ホテル	55. 狩り
21. (旅館 | りょかん)	56. (自然 | しぜん)
22. (民宿 | みんしゅく)	57. 四季
23. の (宿 | やど)	58. (野鳥 | やちょう)
24. (宿 | やど)(案内 | あんない)	59. (紅葉 | こうよう)
25. (温泉 | おんせん)	60. (アプローチ | あぷろーち |approach)
26. (景勝 | けいしょう)	61. (リゾート | りぞーと)
27. (名所 | めいしょ)	62. プロムナード
28. (旧跡 | きゅうせき)	63. (レジャー | れじゃー)
29. (古墳 | こふん)	64. スポーツ
30. (遺跡 | いせき)	65. アミューズメント
31. 跡 ($|\s)	66. (アウトドア |outdoor)
32. (発掘 | はっくつ)	67. (キャンプ | キャンピング |camping)
33. (寺院 | じいん)	68. (散策 | さんさく)
34. 神社	69. (公園 | こうえん |park)
35. 寺 ($|\s)	70. 体育 (施設 | しせつ)

表 A.8: 観光・レジャーの特徴語辞書：その 2(正規表現)

1. (体育館 | たいいくかん)	22. (伝統 | でんとう)(工芸 | こうげい)
2. (ランド | らんど)	23. 塗り?($|\s)
3. (遊 | あそ)(び | ぶ)	24. (市場 |market)
4. (海水 | かいすい)(浴場 | よくじょう)	25. 朝市
5. スキー	26. (山 | やま | 海 | うみ) の (幸 | さち)
6. ゴルフ	27. (味覚 | みかく)
7. (釣り |fishing)	28. (旨 | うま)(い | か)
8. つり情報	29. まいもん
9. 地引き?網	30. (食 | た) べて
10. 登山	31. の (味 | あじ)
11. (海岸 | かいがん)	32. (味 | あじ)(の | わ)
12. (時刻 | じこく)(表 | ひょう)	33. (グルメ | ぐるめ)
13. 自動車道	34. (直通 | ちょくそう)(便 | びん)
14. 高速道路	35. ふるさと (からの)?(便 | たよ)
15. 鉄道	36. の (楽 | たの) しみ (方 | かた)
16. (特産 | とくさん)	37. ぼーと (過 | すご) す
17. (物産 | ぶっさん)	38. (リフレッシュ| りふれっしゅ)
18. (名産 | めいさん)	39. (体験 | たいけん)
19. (名物 | めいぶつ)	40. (見学 | けんがく)
20. (土産 | みやげ)	41. (巡り | めぐり)($|\s)
21. (民芸 | みんげい)(品 | ひん)	42. (観光 | かんこう)

表 A.9: 住民向けの特徴語辞書 (正規表現)

1. (図書館 | としょかん)
2. (情報 | じょうほう)(公開 | こうかい)
3. (災害 | さいがい)
4. (防災 | ぼうさい)
5. 交通 (規制 | きせい | 対策 | たいさく)
6. (避難 | ひなん)(場所 | ばしょ)
7. (消防 | しょうぼう)
8. (catv|cabletelevision)
9. (選挙 | せんきょ)
10. (行政 | ぎょうせい)
11. (議会 | ぎかい)
12. ⟨ 地域タイプ ⟩ 政
13. 施政
14. 憲章
15. 条例
16. 税金
17. 財政
18. (報償 | 奨励) 金
19. (地域 | ちいき)(振興券 | しんこうけん)
20. (2000 年問題 |Y2K)
21. ⟨(地域タイプ | 地域名)⟩ の?組織
22. (課別 | かべつ)
23. (水 | すい)(道 | どう)
24. (相談 | そうだん)
25. (福祉 | ふくし)
26. (介護 | かいご)(保険 | ほけん)
27. (ゴミ | ごみ)
28. (生涯 | しょうがい)(学習 | がくしゅう)
29. (生 | い) きがい
30. ふれあい
31. サークル
32. 保育所
33. パソコン (教室 | きょうしつ)
34. 住まい
35. (暮 | く) らし
36. ⟨(地域タイプ | 地域名)⟩ 民の?(生活 | せいかつ)
37. (便利 | べんり)(帳 | ちょう)
38. 茶の間ガイド
39. (広報 | こうほう)
40. (旬報 | しゅんぽう)
41. ⟨(地域タイプ | 地域名)⟩ 報
42. (病院 | びょういん)
43. (郵便局 | ゆうびんきょく)
44. 公共 (施設 | しせつ)
45. ⟨(地域タイプ | 地域名)⟩ の (施設 | しせつ)
46. ⟨(地域タイプ | 地域名)⟩(民 | みん)(会館 | かいかん)
47. 地区 (会館 | かいかん)
48. (公民 | こうみん)(館 | かん)
49. (健康 | けんこう)
50. (保健 | ほけん)
51. ⟨(地域タイプ | 地域名)⟩(民 | みん) の?(広場 | ひろば)
52. (U|I) ターン
53. ⟨(地域タイプ | 地域名)⟩ からのお (知 | し) らせ
54. ご (連絡 | れんらく)
55. 公共 (団体 | 機関)
56. (官公庁 | かんこうちょう)
57. ⟨(地域タイプ | 地域名)⟩(庁 | ちょう)
58. (役所 | やくしょ)
59. (役場 | やくば)