地域情報ウェブディレクトリの自動編集

大槻洋輔[†] 佐藤理史[††,†††]

本論文では，地域情報ディレクトリを自動編集するシステムを提案する．本システムの主要な処理は，特定の地域に関する情報を提供するウェブサイト（地域サイト）の自動収集と，地域サイト内のウェブページの自動分類である．地域サイトの自動収集では，地域サイトの URL として典型的に用いられる URL パターンを利用して地域サイトのリンク集を発見し，そこから未知の地域サイトの URL を収集する．ウェブページの自動分類では，ウェブページのタイトルやアンカテキストなどに現れるカテゴリ固有表現に基づき，ページを 8 種類のカテゴリに分類する．実験において，本システムは，日本の全地域（3,427 自治体）の 83.2%の地域に対して，地域サイトを 1 つ以上収集することができた．また，ウェブページの自動分類の再現率は 71.4%，適合率は 83.2%であった．

Automated Editing of a Web Directory of Regional Information

Yousuke Ohtsuki[†] and Satoshi Sato[††,†††]

This paper proposes a system that edits a web directory of regional information automatically. Two key technologies are automatic collection of regional web sites and automatic classification of web pages in the collected sites. The former finds existing link collections by using the URL prototype of regional web sites, and collects unknown regional web sites' URLs from these link collections. The latter classifies the pages in the regional web sites into eight categories. The categories of a page are determined by the category-specific expressions that exist in the page title and the anchor texts. In the experiment, the system collected 4,012 regional web sites in total; they cover 83.2 percent of all regions in Japan. The system achieved 71.4 percent recall and 83.2 percent precision in an open test of automatic classification.

1. はじめに

近年，ワールドワイドウェブ（以下，ウェブと略記）が急速に普及し，ウェブを通してさまざまな情報を簡単に入手することが可能になった．これにともない，ある特定の地域に関する情報（以下，地域情報と呼ぶ）も，都道府県や市町村といった自治体の公式ホームページを中心に各種の情報が入手可能となってきている．それらの中には，旅行に役立つ情報や行政サービスの案内など，多くの有用な情報が含まれており，その量と質は，旅行専門誌や広報誌に匹敵する場合も多い．

このような地域情報を見つけ出す主要な方法には，サーチエンジンによるキーワード検索と，ウェブディレクトリを用いた探索の 2 つの方法がある．しかし，これらの方法には，それぞれ問題点が存在する．

サーチエンジンによるキーワード検索の問題点は，検索結果が十分に絞り込まれない形で得られるという点である．ほとんどの場合，検索結果として大量のページ（URL）が得られるが，それらの中には不要なページがかなり含まれているのが普通である．特に，地域情報を探す際には，入力キーワードとしてその地域の名称を選ぶのが普通であるが，このキーワードは非常に多くのページに出現するため，不要なページが大量に検索される傾向が強い．

一方，ウェブディレクトリの問題点は，作成や更新にかなりの労力を必要とする点である．このため，頻繁に更新することができず，すでに存在しな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．

本論文では，後者のウェブディレクトリがかかえる問題点，すなわち，作成や更新の労力を軽減する 1 つの方法として，地域情報ディレクトリを自動編集する

† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology
†† 京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
††† 科学技術振興事業団さきがけ研究 21「情報と知」領域グループ
"Information and Human Activity", PRESTO, JST

システムを提案する．

本システムは，情報源として次の 2 種類のウェブサイトを利用する．

(1) 地域サイト
特定の地域の幅広い情報を提供するサイト．たとえば，自治体の公式ホームページなど．

(2) 特定情報サイト
特定の種類の情報を日本全国の地域に対して提供しているサイト．たとえば，国勢調査ホームページなど．このサイトには，日本全国の各地域の人口と世帯数が掲載されている．

現在，日本には 3,000 以上の自治体が存在し，それらの多くは固有の地域サイトを持つ．これらの地域サイトをすべて手作業で見つけ出すことは，非常に大変な作業になる．本システムでは，この作業を自動化する．

単に，地域サイトを見つけ出すだけでは，地域サイトを地理区分によって分類したリンク集しか作成できない．本ディレクトリでは，地理区分のほかに，8 種類のカテゴリからなる内容区分を導入し，これら 2 種類の区分を用いて地域情報を組織化する．このような組織化を可能とするために，本システムは，地域サイト内のそれぞれのページを 8 種類のカテゴリに自動分類する☆．

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．本システムは，表解析を用いて，これらのサイトから情報を自動抽出する．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ことも行う．

以下では，まず 2 章で，システムが自動生成する地域情報ディレクトリについて述べる．3 章では，システムの概要とその主要な処理について述べる．4 章では実験について述べ，5 章では，検討と関連研究について述べる．最後に，6 章で結論を述べる．

2. 地域情報ディレクトリ

本システムで自動編集される地域情報ディレクトリは，地域表示モードと，カテゴリ表示モードの 2 つの表示モードを持つ．

地域表示モードは，日本全国の 47 都道府県と 3,380 市町村の合計 3,427 自治体☆☆のそれぞれに対して，そ

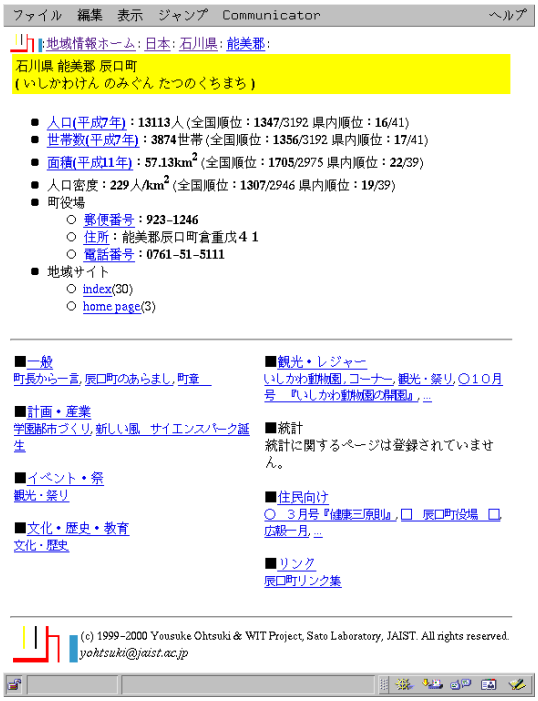

図 1 地域表示モード
Fig. 1 Regional view.

の地域の情報を 1 ページにまとめて表示するモードである．地域表示モードの例を**図 1** に示す．このページは，「石川県辰口町」に対するページであり，ページ上部には，人口，世帯数，面積，人口密度，役所情報，地域サイトへのリンクが表示される．ページ下部には，地域サイトのページを 8 種類のカテゴリ（『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』）に分類した結果が表示される．

カテゴリ表示モードは，ある特定のカテゴリに対するページを，地方や都道府県単位で表示するモードである．このモードでは，カテゴリと，それをより詳細に限定するキーワード，および，対象地方（または都道府県）を指定することができる．表示例を**図 2** に示す．このページは，カテゴリとして『観光・レジャー』，キーワードとして「温泉」，対象地方として「北陸地方」を指定した場合に表示されるページである．このページには，北陸地方のいずれかの地域の観光・レジャーに関するウェブページのうち，キーワード「温泉」を含むページの一覧が表示される．

3. 自動編集システムの概要

上記の地域情報ディレクトリは，本研究で作成したシステムによって自動生成される．そのシステムの構

☆ この分類は排他的分類ではなく，1 つのページが複数のカテゴリに分類される場合もある．

☆☆ 1999 年 6 月 11 日に郵政省ホームページから入手した新郵便番号データベースに基づく．

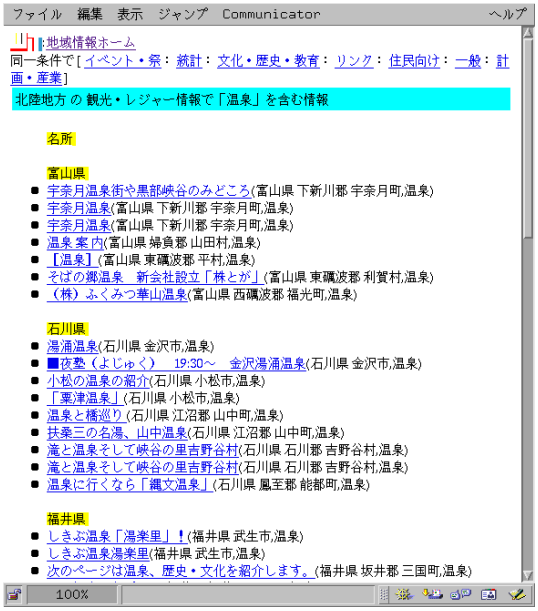

図 2 カテゴリモード
Fig. 2 Category view.

成を**図 3** に示す．本システムは，ディレクトリのコンテンツを作成する**コンテンツ生成モジュール**，作成したコンテンツを格納する**コンテンツデータベース**，ユーザとのインタラクションにより，ディレクトリを表示する**ユーザインタフェース**の 3 つの要素から構成される．

本システムの主要部は，ウェブを通して入手可能な情報からディレクトリのコンテンツを作成するコンテンツ生成モジュールである．このモジュールで行う主要な処理は，地域サイトの自動収集，ウェブページの自動分類，情報抽出・生成の 3 つである．以下の 3 つの節では，これらの処理について説明する．

3.1 地域サイトの自動収集

自治体（都道府県や市町村）の公式サイト☆は，多くの場合，その地方の地域情報を整理した形で提供している．また，公式サイトではないにもかかわらず，ある特定の地域の豊富な情報を提供するサイトも数多く存在する．地域情報ディレクトリの最も重要なコンテンツは，このようなサイト（地域サイト）へのリンクである．このため，できるだけ多くの地域サイトを発見することが必要になる．

好都合なことに，ウェブには，地域サイトの URL を集めたリンク集（地域サイトリンク集）が多数存在する．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の URL を容易に収集することができる．

では，どのようにして地域サイトリンク集を発見すればよいであろうか．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の 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，「地方公共団体ドメイン名」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（自治体）のためのドメイン名であり，以下のルールで作成される☆☆．

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）のいずれかの値を代入する．また，「地域名」と「都道府県名」には，その地域の地名のローマ字表記を代入する．なお，地方公共団体が都道府県の場合は，地域名は存在しない．また，政令指定都市の場合は，「都道府県名」を省略する．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/

すべての地域サイトがこのパターンの URL をとるわけではないが，このパターンの URL をとるサイトは，ここでの収集対象である地域サイトであることはほぼ間違いない．そこで，このようなパターンの URL を多数リンクしているページを見つけることができれば，そのページは地域サイトリンク集である可能性が高い．

こうして地域サイト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 を収集することができる．さらに，ここで収集された URL を利用して，再度リンク集の探索を行えば，前回の探索で見つからなかったリンク集を発見できる可能性もある．

以上の考え方に基づき，次の方法で地域サイトを収集する．

(1) 地方公共団体ドメイン名に対応する URL を，すべての地方公共団体（自治体）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) それぞれの都道府県に対して，以下の処理を行う．

☆ 本論文では，都道府県庁や市役所，町村役場が作成したもの，あるいは，連絡先がそうなっているものを公式サイトと見なす．

☆☆ http://www.nic.ad.jp/jpnic/domain/saisoku-1.html

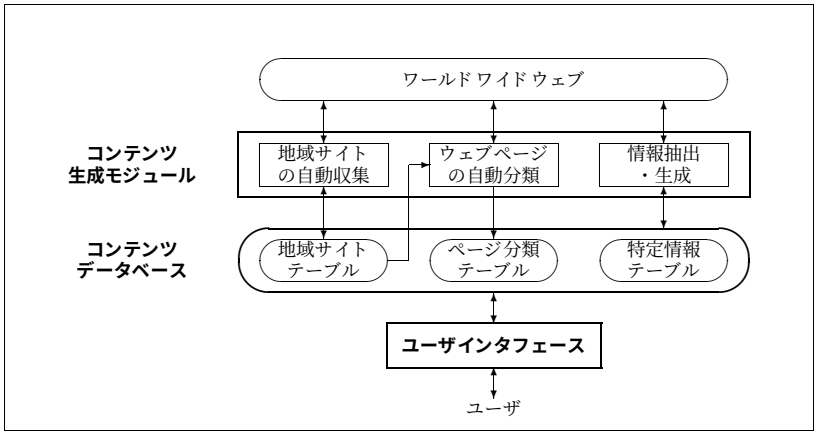

図 3 自動編集システムの構成
Fig. 3 Configuration of the system.

(a) その都道府県に属するすべての地域に対する地域サイト URL を，地域サイトテーブルから取り出す．同一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☆が最も多いものを選ぶ．

(b) 取り出した URL 群の中から，逆リンク数が多い URL を上位 10 個選び，この 10 個の 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情報リンク集とする．

(c) 見つかったリンク集から，アンカテキスト☆☆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．この URL に対応するウェブページが実際に存在する場合，この URL をその地域名に対応する地域サイトとして，地域サイトテーブルに登録する．

この方法のステップ (2) で，都道府県単位に処理を行うのは，地域サイトリンク集の多くが都道府県単位となっているためである．また，ステップ (2) (b) で URL を 10 個に限定するのは，使用しているサーチエンジンの制約による．

地域サイトの収集は，繰り返し実行する．2 回目以降は，上記のステップ (2) のみを実行する．これにより，2 回目以降の収集では，前回の収集で見つかった地域サイト URL が利用されることになる．

☆ 他のページからそのページへのリンク数．
☆☆ アンカタグ（〈a〉と〈/a〉）で囲まれる文字列．

地域サイトテーブルに URL を登録する際には，石田ら[1)]がダングリングリンクと呼ぶ切断リンクの自動メンテナンス処理を行う．この処理は，以下の 2 種類のダングリングリンクに対して行う．

(1) 移動通知ページ
石田らの方法[1)]に準ずる方法で行う．

(2) or ドメイン → ne ドメイン
URL が or ドメインであり，その URL に対応するページが存在しないか，あるいは，そのページ内に ne ドメインへ変更した URL の記述がある場合，対応する ne ドメインの URL に実際にページが存在すれば，or ドメインを ne ドメインに変更する．

3.2 ウェブページの自動分類

ウェブページの自動分類では，地域サイト内のページを内容別に自動分類する．分類カテゴリとして，『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 つのカテゴリを用いる．これらの 8 つのカテゴリは，地域サイトで実際に用いられている分類カテゴリと，Cyber City Case Bank☆☆☆で用いられている分類カテゴリを参考にして決定した．

ページに付与するカテゴリは，そのページを代表する以下の 3 種類のテキスト（判定対象テキスト）に現れるカテゴリ固有語（それぞれのカテゴリに固有な単語や表現）に基づいて決定する．

(1) そのページのタイトル

☆☆☆ 地域情報ディレクトリの 1 つで，内容区分による分類を取り入れている．1998 年 7 月に更新を休止した．URL は http://www.ccci.or.jp/city-cb/であったが，2001 年 3 月の時点では消滅している．

表 1 カテゴリ固有語辞書
Table 1 Dictionary of category-specific expressions.

カテゴリ	語数	カテゴリ固有語（一部）
一般	32	〈地域名〉の概要，〈地域名〉の沿革，国際交流，姉妹都市，〈首長〉のあいさつ，···
計画・産業	22	プラン，工場立地，都市づくり，工業，農業，プロジェクト，分譲，テクノパーク，···
イベント・祭り	20	イベント，行事，開催，歳時記，観賞，まつり，祭り，催し，大会，一般公開，···
文化・歴史・教育	37	文学，民話，著名人，狂言，歴史，伝統，教育，方言，芸術，作品，遺跡，工房，···
観光・レジャー	112	位置，宿泊，名所，特産，寺院，レジャー，美術館，スポーツ，アクセス，公園，···
統計	8	データ，数字で見る，統計，数の推移，指数，指標，面積，人口
住民向け	59	行政，広報，お知らせ，防災，税金，〈地域タイプ〉の組織，図書館，財政，条例，···
リンク	3	リンク，りんく，link

(2) そのページを指すアンカのアンカテキスト
(3) そのページ内の強調テキスト☆1

使用するカテゴリ固有語辞書は，石川県内の 41 地域，54 サイトにあるページを人手で 8 種類のカテゴリに分類し，それらのページのタイトル，アンカテキスト，強調テキストから，カテゴリ固有語を選び出すことによって作成した☆2．カテゴリ固有語辞書の一部を**表 1** に示す．

ページ分類は，次の手順で行う．

(1) アンカテキストとタイトルを調べる．カテゴリ固有語が含まれていた場合，そのカテゴリ固有語に対応したカテゴリをそのページに付与して終了する．なお，異なるカテゴリに属する複数のカテゴリ固有語が含まれる場合は，複数のカテゴリを付与する．
(2) 強調テキストを調べる．カテゴリ固有語が含まれていた場合，そのカテゴリ固有語に対応したカテゴリをそのページに付与して終了する．
(3) 分類不能とする．

なお，分類の対象とするページは，地域サイト内のページのうち，トップページから距離 2 以下のページとする☆3．ただし，そのページのタイトル，アンカテキストに，「新着」「索引」「更新履歴」や該当地域以外の地域名が含まれている場合，そのページからリンクされているページは分類の対象としない．

8 種類のカテゴリのいずれかが付与されたページは，その URL を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．ただし，あるページのカテゴリがその親ページのカテゴリと一致する場合，そのページは登録しない．これにより，それぞれのカテゴリに分類されるページが無意味に増加することを防ぐ．たとえば，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」がカテゴリ『観光・レジャー』に分類された場合を考えよう（**図 4**）．この場合，これら 2 つのページはカテゴリ分類テーブルには登録しない．一方，親ページのカテゴリと異なるカテゴリに分類されたページ（「イベント」）は登録する．

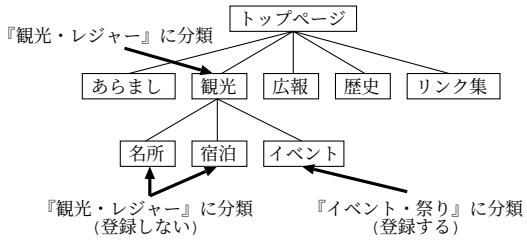

図 4 地域サイトにおけるページ分類
Fig. 4 Page classification in a regional site.

3.3 情報抽出・生成

情報抽出・生成では，特定情報サイトから各地域に対する情報を抽出することと，それらの情報を組み合わせて，新しい情報を生成する処理を行う．

情報抽出では，**表 2** に示す 3 つの特定情報サイトから情報を抽出する☆4．抽出した情報は，属性–属性値の形式で，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．

特定情報サイトの数は限られているので，最悪の場合は，それぞれに対して専用の抽出プログラムを用意すればよい．しかし，上記の 3 つの特定情報サイトは，いずれも情報が表形式で記述されており，そのトップページは，次の 2 種類に分けられる．

- 目次のみのページ：47 都道府県に対するアンカが記述されている．そのリンク先ページに，都道

☆1 そのページで特に強調されている 50 バイト以下の文字列を最大 3 つまで抽出したものを強調テキストとする．多くの場合，そのページ中に存在する大きな見出しが抽出される．
☆2 その語や表現を見ただけで，カテゴリを一意に決定できるような語や表現を選び出し，カテゴリ固有語とし辞書に登録した．その後，予備実験により辞書の調整を行った．なお，固有名詞はカテゴリ固有語として登録しないこととした．
☆3 ハイパーリンクを 1 回たどることを距離 1 とする．
☆4 このうち，平成 7 年度国勢調査は，http://www.stat.go.jp/data/kokusei/1995/08.htm に移動している．

表 2 情報抽出の対象とする特定情報サイト
Table 2 Web sites that the system extracts information from.

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

府県内の各地域に対する情報を記載した表が存在する．

- 都道府県に対する情報ページ：47 都道府県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，都道府県内の各地域の情報を記載した表が存在する．

これらの点を考慮して，以下のような表解析を中心とした方法をプログラム化し，このプログラムによって情報を抽出することとした．

(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[2] 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属性–属性値のリストとして抽出する．
(3) そのページに地域名をアンカ文字列とするアンカが存在した場合は，その URL を抽出して，1 から 3 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 以内の場合に行う．

情報生成では，情報抽出によって得られた数値データを組み合わせて新たな情報を生成する．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
  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から，計算する．
- 人口，世帯数，面積，人口密度の全国，同一都道府県内での順位

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．

4. 実　　験

4.1 地域サイトの自動収集

3.1 節で述べた方法に基づき，地域サイトを収集する実験を行った☆．実験結果を**表 3** に示す．

この表において，URL 作成時とは，地方公共団体ドメイン名に対応する URL を作成した時点を表す．全地域に対して作成した 3,427 個の URL のうち，実際にページが存在したのは 719 件（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て地域サイト URL を抽出した結果，地域サイトを発見できた地域は，2,725 地域（79.5%）に増加した．この処理をもう 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，最終的に，2,852 地域（83.2%）に対して 1 つ以上の地域サイトを見つけることができた．発見総数は，4,012 サイトであり，1 地域あたりの平均サイト数は，1 回目終了時は 1.30，2 回目終了時は 1.41 であった．

表 3 地域サイトの自動収集の実験結果
Table 3 Experimental result: collection of regional sites.

	地域数	サイト数	サイト/地域
URL 作成時	719 (21.0%)	719	1.00
1 回目終了時	2,725 (79.5%)	3,532	1.30
2 回目終了時	2,852 (83.2%)	4,012	1.41

これらの結果より，提案手法によって大量の地域サイトを収集できることが分かった．

4.2 ウェブページの自動分類

3.2 節で述べた方法に基づき，地域サイト内のページを自動分類する実験を行った．実験では，カテゴリ固有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 テストを行った．

実験結果を，判定対象テキスト別に整理したものを**表 4** に示す☆☆．この表から，分類の適合率は，アンカテキストが 90%前後，ページタイトルが約 80%であるのに対し，強調テキストは，60%台にとどまっていることが分かる．

この表で特に注目すべき点は，アンカーテキストとページタイトルの適合率が open テストでもほとんど低下していない点である．この結果から，再現率よりも適合率を重視する必要がある場合には，強調テキストを利用しない方がよいと判断できる．

一方，再現率は，open テストでは，closed テストと比べて大きく低下した．この原因を調べたところ，

☆ 本章で述べる 2 つの実験は，1999 年度下半期に行った．

☆☆ ページ分類の評価は，ページ単位ではなく，カテゴリ単位で行った．すなわち，1 ページに 2 つのカテゴリを割り当てるべき場合は，これを 2 と数えた．

表 4 自動分類の実験結果
Table 4 Experimental result: automatic classification.

判定対象テキスト	closed test					open test				
	Should	Assign	Correct	再現率	適合率	Should	Assign	Correct	再現率	適合率
アンカテキスト (A)	–	634	577	–	91%	–	611	540	–	88%
ページタイトル (T)	–	152	123	–	81%	–	85	68	–	80%
小計 (A+T)	829	786	700	84.4%	89.1%	962	696	608	63.2%	87.4%
強調文字列 (E)	–	43	29	–	67%	–	130	79	–	61%
合計 (A+T+E)	829	829	729	87.9%	87.9%	962	826	687	71.4%	83.2%

Should：割り当てるべきカテゴリの総数，Assign：割り当てたカテゴリ数，Correct：正解

その過半数がカテゴリ固有語の不足であった．すなわち，作成した辞書は，カバレジの点で不十分であり，カテゴリ固有語をさらに追加する必要がある．

5. 検討と関連研究

5.1 検　　討

地域情報の自動編集システムの中心技術は，地域サイトの自動収集とウェブ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，日本全国の 83.2%の地域に対して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，リンク集を発見することができる．

(2) 地域サイトリンク集が，多数存在している．このようなリンク集において，地域サイトへのハイパーリンクのアンカテキストには，ほとんどの場合「地域名」が用いられるため，地域サイト 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[3] や Sato らのリンク集の自動生成[4] でも有効に働くことが報告されている．上記の理由 (1) は，収集しようとしているページの URL のサンプル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，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テキストなどに見られるカテゴリ固有の表現に着目し，簡便な方法でページの分類を実現した．この方法は，高い適合率を達成できることができた．これは次の理由によると考えられる．

(1) ページの内容がアンカテキストやタイトルだけで理解できることが多い．

(2) 地域情報のみを対象としているため，カテゴリ固有語が曖昧性を持つことが少ない．

再現率に関しては，カテゴリ固有語辞書が不十分であったため，open テストでは大きな低下がみられた．closed テストの結果（再現率 87.9%，適合率 87.9%）は，この方法の性能の上限の目安を与えている．我々の当初の目標は，再現率 90%，適合率 90%であり，理想的な辞書を用意できれば，この方法によってそれに近いレベルの精度を達成できることが判明した．残された問題は，カテゴリ固有語辞書の増補によって，open テストの精度（特に再現率）を closed テストの精度に近づけることが可能かどうかである．これについてはさらなる研究が必要であるが，(1) open テストで適合率がそれほど低下していない，(2) 各カテゴリにおいて，再現率と適合率に大きなばらつきがある，という観察結果と，(3) これまでの分類の予備実験と辞書の修正を行ってきた経験，の 3 点から総合的に判断して，辞書の改良の余地は十分に残されていると考えている．

5.2 既存の地域情報ディレクトリとの比較

本システムで自動生成した地域情報ディレクトリと，既存のディレクトリ・リンク集との比較を**表 5** に示す．

一般に，ウェブディレクトリの有用性には多くの要素が複雑に関係しており，それらを比較することはそれほど単純ではない．そこで，ここでは，次の 3 点を取り上げ，比較を行った．

(1) 収録総サイト数：これは，そのディクトリのサイズ（量）を測る 1 つの指標となる．

(2) 収録サイトの種類：これは，そのディレクトリの質を測る 1 つの指標となる．ここでは，公式サイト（official），非公式サイト（unoffical），移動したページや存在しないページへのリンク（out of date）の 3 種類に分類した．一般に公式サイトは有用であることが多く，まず一番にディレクトリに収録すべきサイトである．これに対して，非公式サイトは玉石混交であり，む

表 5 既存のリンク集・ディレクトリとの比較
Table 5 Comparison of web directories.

	Ours	ana	CCCB	Yahoo!
総サイト数	4,012	2,842	2,797	8,212
石川県内				
official	41	41	26	24
unofficial	15	2	4	64
out of date	1	3	28	9
合計	57	46	58	97
分類（内容区分）	Yes	No	Yes	No

Ours：本システムで自動作成したディレクトリ
ana：全国自治体リンク集（ana 版）
(http://www.nsknet.or.jp/˜ana/jiti/)
CCCB：Cyber City Case Bank
(http://www.ccci.or.jp/city-cb/)
Yahoo!：YAHOO!JAPAN「地域情報 → 都道府県」
(http://dir.yahoo.co.jp/Regional/Prefectures/)

やみやたらに収録すればいいわけではない．存在しないページへのリンクの存在はディレクトリの質を大きく低下させるため，少ない方が好ましい．

(3) 内容区分による分類の有無：地域区分以外のアクセス方法が用意されていることが好ましい．

収録サイト数が最も多いのは Yahoo! である．しかし，Yahoo! は，非公式サイトの数が多く，公式サイトの収録もれが多い．公式サイトを最も網羅しているのは，全国自治体リンク集（ana）であるが，このリンク集は，地理区分による分類だけであり，内容区分による分類を提供していない．Cyber City Case Bank（CCCB）は，内容区分の分類を提供しているが，更新が休止されており，移動したページや存在しないページへのリンク（out of date）がかなり含まれている．

自動生成したディレクトリは，収録している公式サイト数では ana と遜色がなく，かつ，CCCB と同様に内容区分の分類を提供している．分類の精度は人間並みとはいかないが，総合的に考えて，上記の 3 点の比較においては，本ディレクトリは既存のディレクトリ・リンク集にほぼ匹敵するレベルに達していると見なすことができる．

今回作成した自動編集システムでは，著作権の問題から，地図や画像など情報をディレクトリに含めることを見送ったが，これらをディレクトリに含めることは技術的には何の問題もない．このようなコンテンツの集積をさらに行えば，人手で作成したディレクトリを越えるものを作成できる可能性もある．

5.3 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[3] 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（たとえば，“cheese”）に対して，それに関する少数の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）からなる．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．

Sato ら[4] 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（たとえば,「水族館」）からそのカテゴリに属するインスタンス名（たとえば,「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（地方公共団体名）が分かっているため，このような方法をとる必要がない．

本研究以外に，ウェブ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[5] 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，たとえば，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，地域情報を組織化しディレクトリとして提供することを目的としている．

6. おわりに

本論文では，地域情報ディレクトリを自動編集するシステムを提案し，それを実現する方法について述べた．本システムは，次の特徴を持つ．

- 地域サイトと特定情報サイトの 2 種類の情報源を利用する．
- 地域サイトを自動的に発見し，収集する．
- 収集した情報を地域別に整理するだけでなく，内容別に自動分類する．

地域サイトの自動収集では，4,012 の地域サイトを発見することができた．これは，全国 3,427 の地域のうち，全体の 83.2%の 2,852 地域をカバーしており，実用レベルに達しているといえる．一方，ページの自動分類は，再現率 71.4%，適合率 83.2%であり，実用

レベルまであと一歩というところである．

参 考 文 献

1) 石田，谷川，宮下：WWW におけるダングリングリンクの自動メンテナンス，情報処理学会第 59 回全国大会，Vol.3, pp.85–86 (1999).
2) 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，情報処理学会研究報告 ICS-119-24，pp.173–180 (2000).
3) Members of the Clever Project: Hypersearching the Web, *Scientific American*, Vol.280, No.6, pp.54–60 (1999).
4) Sato, S. and Sato, M.: Toward Automatic Generation of Web Directories, *Proc. International Symposium on Digital Libraries 1999* (*ISDL'99*), pp.127–134 (1999).
5) 三浦，高橋，横路，島：位置指向の情報統合—モーバイルインフォサーチ 2 実験，情報処理学会第 57 回全国大会，pp.637–638 (1998).

(平成 12 年 8 月 7 日受付)
(平成 13 年 6 月 19 日採録)

大槻 洋輔（正会員）

1998 年愛知工業大学経営工学科卒業．2000 年北陸先端科学技術大学院大学情報科学研究科修士課程修了．同年，三洋電機株式会社に入社．

佐藤 理史（正会員）

1983 年京都大学工学部電気工学科第二学科卒業．1988 年同大学院博士課程研究指導認定退学．京都大学工学部助手，北陸先端科学技術大学院大学情報科学研究科助教授を経て，2000 年より京都大学大学院情報学研究科助教授．1997 年より 2000 年まで科学技術振興事業団研究員を兼任．京都大学博士（工学）．自然言語処理，機械学習，情報の自動編集等の研究に従事．言語処理学会，日本認知科学会，AAAI，ACL 各会員．著書：『自然言語処理』（共著，岩波書店，1996),『アナロジーによる機械翻訳』（共立出版，1997),『言語情報処理』（共著，岩波書店，1998）等．