# 地域情報を対象とした ウェブディレクトリの自動生成

大槻 洋輔

北陸先端科学技術大学院大学 情報科学研究科

2000 年 2 月 15 日

**キーワード**: ウェブディレクトリ、自動編集、地域情報、ワールドワイドウェブ、情報抽出.

近年, インターネットが急速に広まり, ワールドワイドウェブ (WWW) 上には, 自治体の公式ホームページなど, さまざまな地域情報が豊富に存在する. 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.

これらの問題点を解決するため, 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.

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.

本システムは, 次の 3 つの要素から構成される.

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**
   ユーザの要求する表示モードに従って, 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.

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.

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, 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, 都道府県別面積サイトなどがこれに該当する.

地域情報編集モジュールで行なう処理は, 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.

**地域サイトの自動収集**は, 本システムが対象としている 3427 自治体に対する地域サイトを自動的に収集する. この実現には, WWW 上に多数存在する, 地域サイトのリンク集を利用する. 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. また, 収集した URL をもとに, 収集処理を繰り返し行なうことで, より多くの地域サイトを収集する. 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.

**地域情報ページの自動分類**は, カテゴリモードを実現するため, 地域サイト内のページを, 「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 つのカテゴリに自動的に分類する. 地域情報ページの自動分類では, まず, それぞれのカテゴリに特有の単語や表現を特徴語としてまとめておく.

そして, あるカテゴリに対する特徴語が, 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場合, そのカテゴリに特徴語が現れたページを分類する. 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. 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, いくつか改善する余地がある.

一方, 特定情報サイトを情報源とした地域情報編集処理は, 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, 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. 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.