

## I-11 WWW 自動探索による電子地図の属性情報自動抽出システムの研究開発

Fundamental Research of System for Extracting Attributes of Digital Map  
by Automatic Search on WWW物部寛太郎<sup>1</sup>・田中成典<sup>2</sup>・古田均<sup>2</sup>・加藤佑一<sup>3</sup>・野中広茂<sup>3</sup>

Monobe Kantaro, Tanaka Shigenori, Furuta Hitoshi, Kato Yuichi, and Nonaka Hiroshige

**抄録:** 近年、情報処理技術の発展に伴い、空間情報を利用したサービスをパソコンや携帯電話を通して誰もが容易に利用できるようになった。しかし、空間情報サービスの基盤となる電子地図データに含まれる属性情報の不足や更新頻度の低さが問題になっている。それらの問題が生じるのは、電子地図データの作成および維持管理のためのコストや労力の高さが原因であると考えられる。そのため、自動的に属性情報を取得するシステムの開発が望まれている。そこで、本研究では、Web サイトを限定せず、WWW を探索して属性情報を自動的に収集し、電子地図データの属性情報の作成を行うシステムの研究開発を目指す。さらに、収集した属性情報を用いて、自然言語をキーワードとして空間情報を検索するシステムの研究開発も行う。

**Abstract:** A service for providing spatial information is increasing recently. This service is easily used by a personal computer or a mobile phone. It is important that a digital map has a lot of attributes on the spatial service. However, making attributes of digital maps requires a great deal of time and money. Therefore, it is difficult to define attributes of spatial data and to produce these values. The purpose of the present research is to develop a system that can produce and update automatically attributes of the digital map by searching them on web pages. Moreover, a function for searching spatial information with the keyword of natural language is implemented.

**キーワード:** GIS, Web, 電子地図, 属性, 自然言語処理

**Keywords:** GIS, Web, Digital Map, Attribute, Natural-Language Processing

## 1. まえがき

近年、情報技術の発展に伴い、空間情報の重要性が非常に高まっている。誰もが簡単に位置情報を取得できるようになり、空間情報は、我々の生活には欠かせないものになりつつある。最近では、空間情報を用いたサービスの提供が増加している。人々は、それらのサービスをパソコンや携帯電話を通して利用することができる。それらのサービスをさらに発展させるためには、空間情報の整備が重要である。

空間情報は、幾何情報と属性情報によって構成される。幾何情報に関しては、情報処理技術の発展に伴い、3Dで表現することが可能になった。3Dによって、現実世界に限りなく近い電子地図が構築されている。しかし、属性情報の整備は未だ進んでいない。例えば、国土地理院刊行の数値地図やWeb上で提供される電子地図は、建物名、住所や電話番号などの最低限の属性情報のみを保持している場合が多い。

このように属性情報が不足している現状では、今後さらにニーズが高まるであろう空間情報サービス

に対して、十分な情報を提供することができない。例えば、「淀川に架かるアーチ橋」を検索する場合、電子地図上の地物に、「淀川」や「アーチ橋」のような詳細な属性情報が保持されていなければ、検索は不可能である。また、GISを防災、都市計画やバリアフリーなどに利用する場合にも、詳細な属性情報は必要不可欠である。

属性情報には、常に情報の最新性が求められる。正確な空間情報サービスを行うためには、空間情報が常に現実世界と等しい状態に保たれている必要がある。そのためには、属性情報の頻繁な更新と整備が欠かせない。現状では、属性情報の整備には多大なコストと労力が必要<sup>2)</sup>となる。そのため、属性情報を容易に整備するためのシステムの開発が求められている。そこで、本研究では、電子地図の属性情報の自動収集の実現を目指す。

## 2. 研究の目的

本研究では、電子地図の属性情報の元データとし

1: 学生会員 情修 関西大学大学院 総合情報学研究科

(〒569-1095 大阪府高槻市霊仙寺町2-1-1, Tel:072-690-2153, E-mail: mkkkm@aurora.dti.ne.jp)

2: 正会員 工博 関西大学 教授 総合情報学部

3: 非会員 情学 関西大学大学院 総合情報学研究科

てWWW上の情報に着目した。現在、我が国では、8,590万のWebページが存在<sup>3)</sup>している。そのWebページ内に含まれる位置情報と電子地図上の地物を対応させることにより、WWW上の情報を属性情報として付加することが可能になると考えられる。

既存の研究としては、相良ら<sup>4)</sup>によって、Web上に存在するジオリファレンス情報を取り出し、その情報を位置座標に変換して、空間情報サービスに有効利用するための研究が行われている。実システムとしては、空間情報抽出システムや空間情報サーチエンジンを開発している。さらに、XML表現を利用した空間タグを埋め込むことを提案し、その有用性の評価も行っている。中嶋ら<sup>5)</sup>は、Webサービスによる地図検索システムを構築している。SOAPメッセージングによるサービスが行われており、地図情報の配信などを可能にしている。斉藤ら<sup>6)</sup>は、Semantic Webを利用した地理情報検索システムに関する研究を行っている。久保ら<sup>7)</sup>は、空間情報と時系列情報を統合したGISモデルシステムを開発している。

しかし、これらの研究では、属性情報の自動抽出までは行われていない。また、対応しているWebサイトが限定されているなどの制約もある。さらに、位置情報が不完全でも情報を収集するため、情報の信頼性が低いという問題点もある。

そこで、本研究では、WWWを自動探索することで、電子地図の属性情報の作成を支援するシステムの研究開発を目指す。さらに収集した属性情報を利用し

て、自然言語から空間情報を検索する機能も実装する。本システムの構想を図-1に示す。本システムでは、最初にWebページから属性情報の抽出を行う。続いて、その中の住所情報を用いてアドレスマッチングを行い、電子地図上の地物に属性情報の付加を行う。また、自然言語による空間情報の検索も実現する。

### 3. システムの詳細

本研究で開発するシステムは、WWWを探索して自動的に属性情報を収集するための属性情報自動収集システムと、自然言語をキーワードとした空間情報の検索を行うための自然言語による空間情報検索システムから構成される。本章では、これら2つのサブシステムの詳細について説明する。

#### (1) 属性情報自動収集サブシステム

本サブシステムは、WWWを探索することによって、自動的に電子地図の属性情報を収集することを可能にする。本サブシステムは、a) WWW自動探索機能、b) HTML解析機能、c) 位置情報抽出機能、d) 属性情報抽出機能、e) 座標情報取得機能、f) 属性情報出力機能の6つの機能により実現する。本サブシステムの構成について図-2に示す。以下に各機能の詳細を説明する。

#### a) WWW自動探索機能

本機能では、Webページ内のリンク情報を辿ることによってWWWの探索を行いWebページを収集する。本機能の流れを図-3に示す。取得する対象は、HTML形

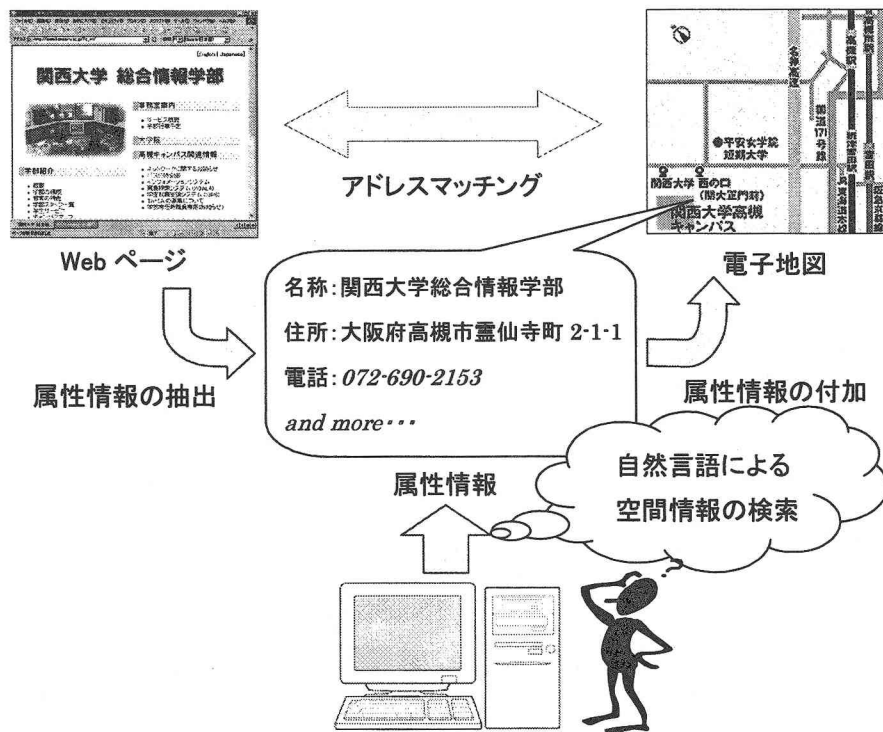


図-1 システムの構想

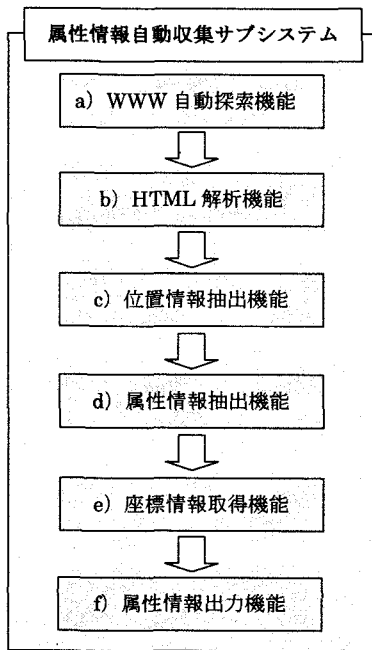


図-2 属性情報自動収集サブシステムの流れ

式のファイルとする。それ以外の EXE ファイルや PDF ファイル等は、Web ページ内にリンクされていても取得しない。その理由としては、例えば、EXE ファイルは住所情報を持っている可能性が低く、PDF ファイルは情報を解析するために大量の時間がかかる、などがあげられる。また、リンク先の URL が相対パスの場合も Web ページの取得は行わない。これは、同じサイト内を繰り返し探索することによる、情報の取得効率の低下を防ぐためである。本システムの自動探索の流れを以下に示す。

- 1) URL をデータベースから取得する。
- 2) 取得した URL の Web ページに接続する。
- 3) 接続した Web ページのファイルを取得する。

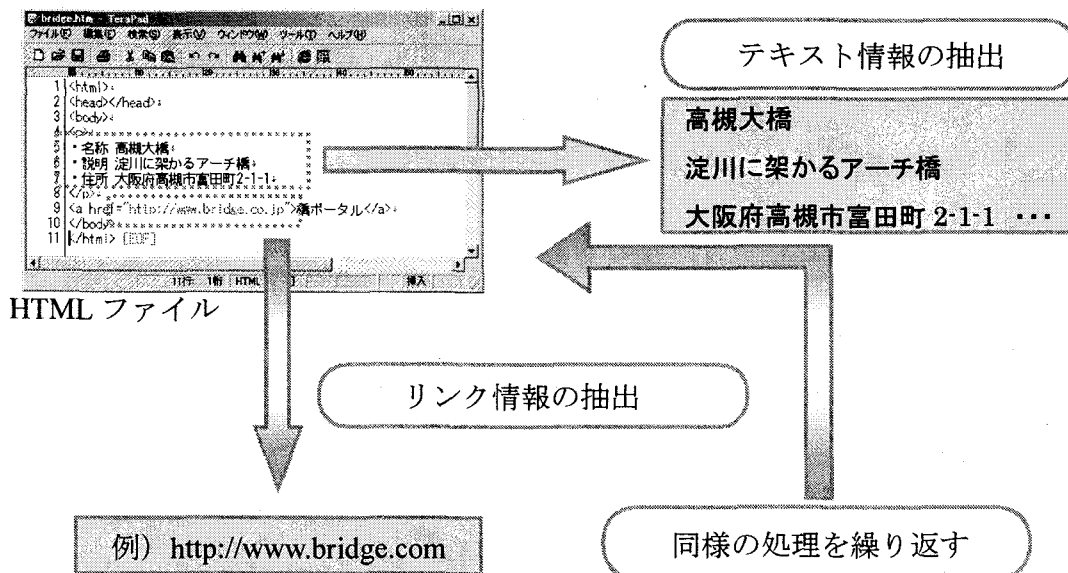


図-3 WWW 自動探索機能の流れ

- 4) Web ページ内のリンク情報を取得する。
- 5) リンク情報をデータベースに保存する。
- 6) 処理 1) に戻り、繰り返す。

本機能では、基点となる Web ページの URL を事前にデータベースに登録しておく必要がある。基点となる Web ページによって、情報の収集結果は大きく異なる。そのため、最初に接続する Web ページの選択は非常に重要である。本研究では、以下のような基準で基点となる Web ページを選択した。

- ・ Web ページ内でリンク情報を多く持つ。
- ・ リンク先が絶対パスである。
- ・ 地名や場所に関連性がある Web ページである。
- ・ 観光サイトやグルメサイトなどのテーマを持つ Web ページである。
- ・ 日本語のページである。

以上の機能により、WWW の自動探索を実現する。

**b) HTML 解析機能**

WWW の自動探索によって収集された Web ページを HTML パーサ技術によって解析する。HTML パーサは、HTML 文書を読み込んで解析するためのソフトウェアである。解析によって、Web ページ内のリンクタグから URL を取得する。さらに、Web ページ内のタグ情報とイメージ情報を除いたテキスト情報を取得する。

**c) 位置情報抽出機能**

本機能では、HTML 解析によって取得したテキスト情報から住所情報を抽出する。住所情報の抽出には、形態素解析を用いた。ここでは Java 用形態素解析システム「Sen」を用いる。形態素解析によって分割された単語の品詞が「地域」であった場合、住所情報として抽出(図-4)する。本システムでは、1つの Web ページに1つの住所情報を持つ場合のみ、住所情報を抽出する。

形態素解析では、位置情報を不完全な形で抽出することがある。例えば、「大阪府」や「高槻市」のみの住所情報の場合は、正確な座標情報に変換することができないため意味を持たない。そのため、住所情報としては「大阪府高槻市霊仙寺 2-1-1」という完全な住所情報として抽出する必要がある。そこで、品詞の並び方をパターン化することによって、完全な情報を抽出するという方法がある。しかし、形態素解析でパターン化して住所情報を抽出するには限界があり、誤った住所情報を取得する可能性がある。そこで、この問題を解決するために、正しい住所情報を記した住所辞書を利用する。本研究では、日本郵政公社が提供している「住所の郵便番号のダウンロードサービス」を利用して、住所辞書を作成する。形態素解析で抽出した住所情報が住所辞書に存在するかを確認することで、誤った住所情報を取り除くことができる。それによって、住所情報抽出の精度を向上させることができる。

**d) 属性情報抽出機能**

本機能では、位置情報の抽出と同様に形態素解析を用いて、属性情報を抽出(図-4)する。形態素解析によって分割された形態素の中から、「名詞」、「形容詞」や「動詞」について取得を行う。これらの情報を取得することで、従来の電子地図には存在しなかった「美しい」などの情報を属性情報として持たせることができる。

**e) 座標情報取得機能**

本機能では、位置情報取得機能によって取得された住所情報から座標情報を取得する。座標情報の取得は、アドレスマッチングにより実現(図-5)する。アドレスマッチングは、東京大学空間情報科学研究センターによって提供されているCSVアドレスマッチングサービス<sup>8)</sup>を利用する。本サービスを利用することによって、住所情報から緯度・経度の座標情報を取得することができる。電子地図上の全ての地物は座標情報を持っているため、座標情報を用いることによって、地物と属性情報をリンク(図-6)させることが可能になる。

**f) 属性情報出力機能**

本機能では、収集した属性情報を URL, 住所情報, 座標情報と属性情報に分類して XML 形式で出力する。XML 形式での出力は、XML-DOM 技術を利用して実現する。XML 形式で出力を行うのは、プログラム上での処理が容易で、GIS による読み込みや加工も可能になるためである。

取得した情報は、データベースによって管理する。データベースシステムには「MySQL」を用いる。MySQL は、リレーショナル・データベースの管理が可能で、大量のデータを高速に検索できるという特徴がある。XML 形式で収集した情報の中で、URL と住所情報を出力した例を図-7 に示す。

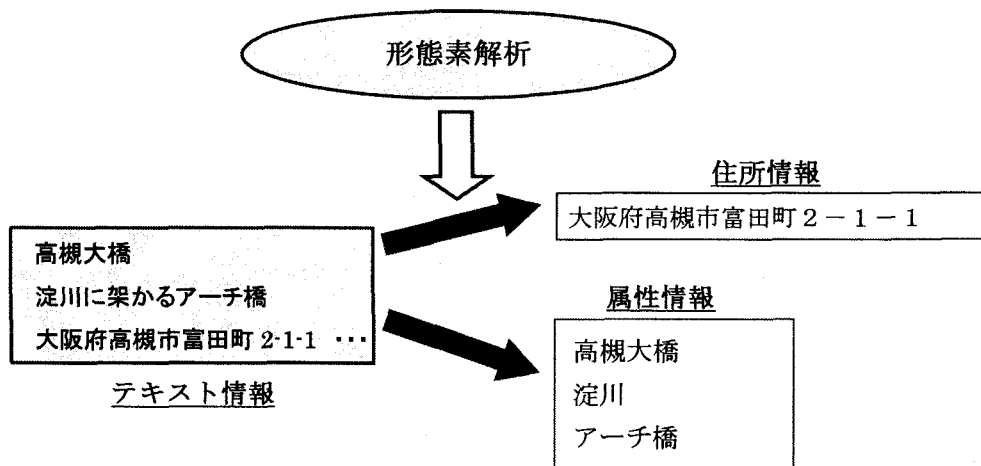


図-4 形態素解析による住所情報と属性情報の抽出

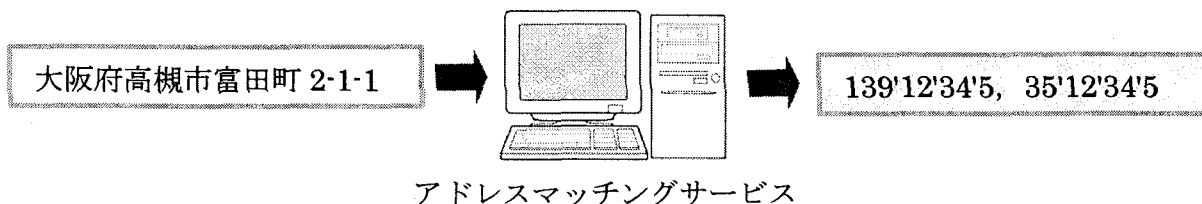


図-5 アドレスマッチングによる座標情報の取得

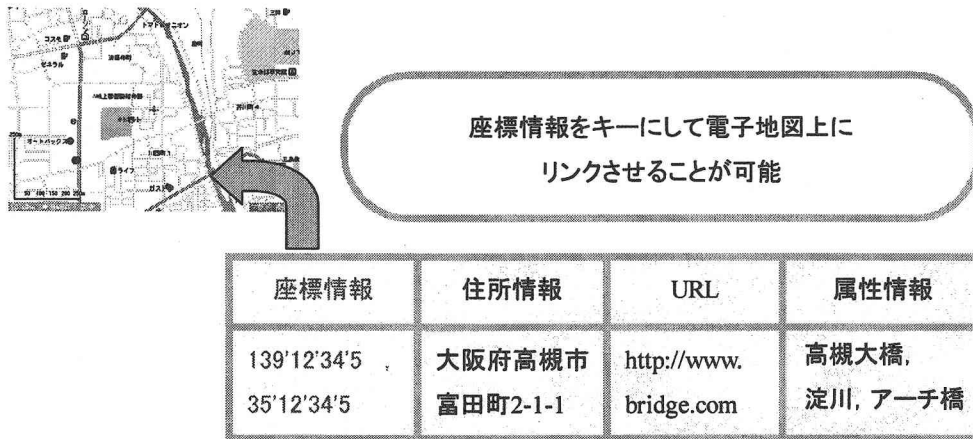


図-6 座標情報による電子地図上の地物と属性情報のリンク

(2) 空間情報検索サブシステム

本サブシステムは、自然言語をキーワードとした空間情報の検索を可能にする。本サブシステムは、a) キーワード入力機能、b) 検索エリア指定機能、c) 地物検索機能、d) 検索結果表示機能の4つの機能により実現する。本サブシステムの構成について図-8に示す。以下に各機能の詳細を説明する。

a) キーワード入力機能

利用者の興味を反映させるため、自然言語によるキーワードの入力を可能にする。自然言語を検索キーワードとすることによって、「楽しい」や「癒し」といった言葉から、その言葉にふさわしい場所を検索することが可能になる。

b) 検索エリア指定機能

本機能では、利用者が検索するエリアを指定する。本システムでは、あらかじめ検索するエリアの電子地図ファイルを選択することによって、エリアを指定する。本システムでは、電子地図として、国土地理院刊行の数値地図 25000 を用いた。電子地図のファイルは、あらかじめ任意のディレクトリに保存しておく。

c) 地物検索機能

本機能では、利用者が入力したキーワードに関連

のある地物の検索を可能にする。自然言語によるキーワードと属性情報自動収集システムによって得られた属性情報のマッチングを行う。この処理によって、利用者が入力したキーワードに適した地物を検索することが可能になる。

d) 検索結果表示機能

本機能では、地物の検索結果を地図上に表示する。利用者が入力した自然言語による検索キーワードに適した地物を分かりやすく表現するために、アイコンを用いて地図上で強調表示する。表示されたアイコンをマウスで選択することによって、その地物の属性情報を取得した Web ページを閲覧することも可能である。検索結果の表示例を図-9に示す。

4. 実証実験

本研究で開発したシステムの有用性を検証するため、実証実験を行った。まず、属性情報抽出の精度を測定するための「属性情報抽出実験」、次に、抽出した位置情報から座標情報を抽出できるかを確認するための「座標情報取得実験」、また、自然言語による空間情報検索の精度を測定するための「空間

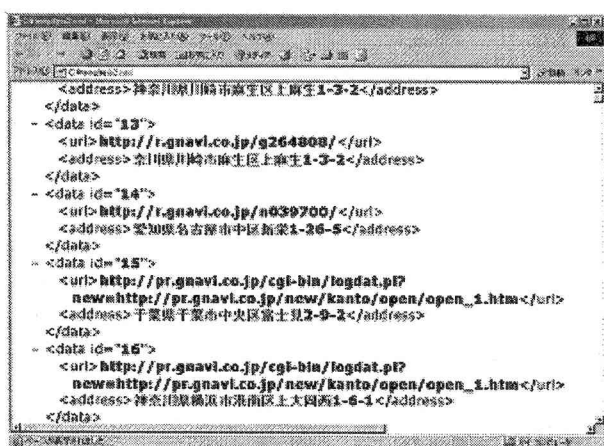


図-7 XML形式で属性情報を出力した例

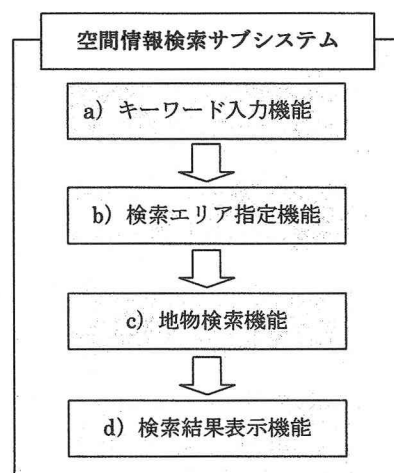


図-8 空間情報検索サブシステムの流れ

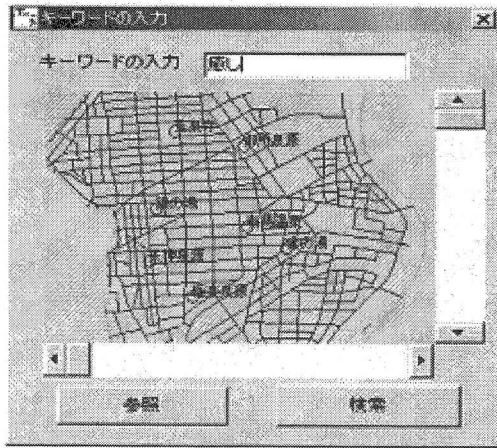


図-9 検索結果の地図上への表示例

情報検索実験」，そして，自然言語による空間情報検索を行った場合の利用者の満足度を調査する「利用者満足度調査実験」を行った。以下に，各実験の詳細を示す。

(1) 属性情報抽出実験

a) 実験方法

本実験では，属性情報自動収集システムを用いて，属性情報の収集を行った。基点となる Web ページは，「ぐるなび (http://www.gnavi.co.jp)」，「旅の窓口 (http://www.mytrip.net)」，「web an (http://weban.engokai.co.jp)」に設定した。これらの Web ページを基点に設定した理由は，住所情報とリンク情報が豊富に存在するためである。今回の実験では，リンクの探索時間は3時間に設定して，WWWの探索を行った。探索終了後，収集した Web ページの中に含まれる住所情報の数を計測した。

b) 実験結果

計測した結果，「ぐるなび」を基点とした場合，WWW自動探索で収集した Web ページは約1万ページであった。収集した Web ページから住所情報を抽出できたのは1,717ページあった。実際に住所情報が

存在する Web ページの確認を行った結果，住所情報を保持するページは，約1,800ページあった。その他の基点から探索を行った場合の結果も含めて，システムの実験結果を表-1に示す。

c) 考察

実験結果より，「ぐるなび」の場合 95.4%，「旅の窓口」の場合 74.3%，「web an」の場合 83.3%の精度で住所情報を取得することができた。本システムでは，住所情報と同時に属性情報も取得してデータベースに保存されるため，位置情報取得の精度は属性情報取得の精度と同じ値となる。したがって，本実験により，平均して 84.3%の精度で属性情報を取得できることが明らかになった。実験結果より，本システムが電子地図データの属性情報の自動作成に有効であることが実証された。

(2) 座標情報取得実験

a) 実験方法

CSV アドレスマッチングサービスを用いた座標情報の取得を行った。CSV アドレスマッチングサービスでは，座標の精度は，番地レベルあるいは町字レベルで結果が返される。アドレスマッチングは，属性情報抽出実験で取得した全4,289件の住所情報に対して行った。

b) 実験結果

アドレスマッチングの結果，表-2に示すように，4,289件中3,269件の住所情報を番地レベルで座標情報に変換することができた。逆に，274件は座標情報を取得することができなかった。

c) 考察

実験結果より，アドレスマッチングによって76.2%の住所情報から番地レベルの精度で座標情報を取得することができた。番地レベルの精度の座標情報があれば，電子地図との正確なリンクがほぼ可能になるため，大半の属性情報は電子地図上で利用可能となることが分かった。ただし，町字レベルでは正確なリンクは不可能である。実験の結果，23.8%

表-1 属性情報抽出実験結果

	ぐるなび	旅の窓口	web an
収集した Web ページ	10,025 件	12,890 件	15,130 件
住所情報を抽出できたページ	1,717 件	1,115 件	1,457 件
実際に住所情報が存在するページ	1,804 件	1,478 件	1,750 件
住所情報抽出の精度	95.2%	75.4%	83.3%

表-2 座標情報取得実験結果

座標精度	件数	割合
番地レベル	3,269 件	76.2%
町字レベル	746 件	17.4%
取得できず	274 件	6.4%
計	4,289 件	100.0%

もの属性情報が電子地図とリンクできないという状態であるため、今後は自然言語処理の精度を向上させて、より多くの属性情報をリンクさせることが必要になると考えられる。

**(3) 空間情報検索実験**

**a) 実験方法**

本実験では、本システムと既存のシステムとの比較を行った。本システムの有効性を確かめるため、8項目の自然言語で比較を行った。本システムと同じ条件で検索を行うため、MapFan では、フリーワード検索、Mapion では、キーワード検索を利用した。また、検索結果の信頼性が保証されていないため、飲食店のポータルサイト「ぐるなび」を利用して、検索キーワードと地物名との関係を確認した。本実験では、「曖昧な自然言語」と「場所を示す言葉」を利用して、Mapion、MapFan と本システムの比較を行った。

曖昧な自然言語をキーワードとした検索では、普段使用する言葉で、特に曖昧な言葉、つまり地物を特定できない言葉を「曖昧な自然言語」として定義する。「曖昧な自然言語」は、個人の趣味嗜好によって解釈が異なり、1つの概念として捉えにくい。例えば、「癒し」、「楽しい」、「ゆったりした」、「急な」、「障害の少ない」などの言葉は、人によって捉え方が異なり、曖昧である。しかし、人々はこのような曖昧な言葉を思い浮かべて、その後、具体的な地物を決定することが多い。本実験では、「曖昧な自然言語」からの地物を検索し、地物の抽出結果を検証する。

場所を示す言葉をキーワードとした検索では、

我々が普段使用する言葉で、曖昧性が少なく、間接的に場所を示す言葉を「場所を示す言葉」として定義する。「場所を示す言葉」は、地物の検索で、場所を直接的に特定することはできないが、ある程度、場所を絞ることのできる言葉である。例えば、「小学校」、「遊園地」、「温泉」などは、場所を1つに特定することができないが、具体的な地物を候補として挙げることができる。本実験では、「場所を示す言葉」からの地物を検索し、地物の抽出結果を検証する。

**b) 実験結果**

「曖昧な自然言語」による検索は、表-3に示すような結果が得られた。キーワード「癒し」においては、Mapion と MapFan がそれぞれ1件と2件の結果を得たことに対して、本システムでは、25件の結果を得ることができた。残り3つのキーワードに関しても、同様に本システムの検索結果の方が良い結果を得ることができた。

「場所を示す言葉」による検索は、表-4に示すような結果が得られた。キーワード「温泉」においては、Mapion と MapFan がそれぞれ51件の結果を得たことに対して、本システムでは、21件の結果を得ることができた。その他のキーワードに関しても、本システムの検索結果の方が少なかったが、キーワード「本屋」に関しては逆の結果が得られた。

**c) 考察**

本実験では、自然言語による地理情報の検索において、「曖昧な自然言語」と「場所を示す言葉」に分けることにより、本研究の有用性を検証した。「曖昧な自然言語」では、既存システムより多くの地物

**表-3 「曖昧な自然言語」による検索結果**

	バリアフリー	楽しい	癒し	知的
本システム	11件	41件	25件	4件
Mapion	0件	0件	1件	0件
MapFan	0件	6件	2件	0件

**表-4 「場所を示す言葉」による検索結果**

	小学校	遊園地	温泉	本屋
本システム	30件	6件	21件	8件
Mapion	46件	8件	51件	1件
MapFan	0件	0件	51件	3件

**表-5 利用者満足度のアンケート結果**

満足度	件数	割合
5	8件	20.0%
4	11件	33.3%
3	7件	23.3%
2	3件	13.3%
1	2件	10.0%

名を検索することができた。しかし、「場所を示す言葉」では、本システムより既存システムの方が多くの地物名を検索できた。

「場所を示す言葉」より「曖昧な自然言語」の方が、多くの地物名を検索できた理由としては、既存のシステムでは、地物を特定することの容易な言葉をカテゴリ分けしており、そのカテゴリに適する地物が検索された場合、検索結果として表示されるためである。しかし、カテゴリ分けされていない言葉は、地物名と一致した場合や、地物の注釈に存在する場合に検索結果として表示されるのみである。そのため、本システムでは、「曖昧な自然言語」においてより多くの地物名を抽出することができた。

#### (4) 利用者満足度調査実験

##### a) 実験方法

本実験では、開発したシステムを用いて自然言語による空間情報検索を行った。被験者は大学生 30 名である。実験後のアンケート結果により利用者の満足度の調査を行った。

##### b) 実験結果

実験の結果、表-5 に示すアンケート結果が得られた。満足度の平均は 3.77 となった。アンケートの際に寄せられた意見によると、肯定的なものでは、「頭でイメージした言葉にあった場所を探すことができるので、楽しく検索することができる」、「自分が今まで知らなかった新しいスポットを検索できそうである」といった意見が寄せられた。否定的な意見では、「自分がイメージした言葉にあった場所が本当に検索されているのか分からない」、「検索結果が少ない」、逆に「検索結果が多すぎる」という意見が寄せられた。

##### c) 考察

満足度は平均を超えているため、概ね本システムによる検索の効果があったと考えられる。しかし、逆に満足度の低い利用者もいるため、今後はそのような利用者のニーズを考慮したシステムの開発を行う必要があると考えられる。

アンケートの際に寄せられた意見によると、自然言語による検索に対して新鮮な印象を受けている被験者がおり、30 件中 19 件という過半数の回答が肯定的な意見であった。自然言語による空間情報処理が評価されたと考えられる。

#### 5. おわりに

本研究では、WWW を自動探索することによって、WWW 上の情報を電子地図の属性情報として利用するためのシステムの開発を行った。本研究で開発した属性情報自動収集システムと空間情報検索システムを用いることによって、

- ・ WWW の自動探索による属性情報の取得
- ・ 形態素解析による HTML からの住所情報と

#### 属性情報の抽出

- ・ XML による属性情報の整備
- ・ 自然言語による空間情報検索

を実現した。また、実証実験より、本システムの有効性を確認することができた。

以上より、本研究によって、電子地図の属性情報作成のコストと労力を削減することができるのではないかと考えられる。さらに、電子地図が詳細な属性情報を持つことによって、様々な土木分野に役立てることができると考える。例えば、築年数や建物種別などの属性は防災分野に、建物の階数や外観などの属性は都市計画に、屋内設備の属性はバリアフリーに役立てることができると考えられる。

**謝辞:** 本研究を遂行するに当たり、文部科学省の私立大学学術研究高度化推進事業における、オープン・リサーチ・センター整備事業による補助を受けた。

#### 参考文献

- 1) Roman Krzanowski, Jonathan Raper: Spatial Evolutionary Modeling, Oxford University Press, 2001.4.
- 2) Brandon Plewe: GIS Online, Thomson Learning, 1997.8.
- 3) 総務省:平成 16 年版 情報通信白書,ぎょうせい, 2004 年 7 月.
- 4) 相良毅,有川正俊,坂内正夫:ジオリファレンス情報を用いた空間情報抽出システム,情報処理学会論文誌:データベース,Vol.41,No.SIG6(TOD7),pp.69-80, 2000 年 10 月.
- 5) 中嶋卓雄,大坪聡志,山名大樹,富松篤典:Web サービスによる地図検索システムの構築,地理情報システム学会講演論文集,Vol.12,pp.383-386,2003 年 9 月.
- 6) 斉藤亮,田中文基,金井理,岸浪健史:Semantic Webを利用した地理情報検索システムに関する研究,地理情報システム学会講演論文集,Vol.11,pp.293-296,2002 年 9 月.
- 7) 久保紀重,飯村威,飯田剛輔,平井政二,大伴真吾:空間情報と時系列情報を統合した GIS モデルシステム開発について,APA,日本測量調査技術協会, No.75-8,pp.67-74,2000 年 3 月.
- 8) 相良毅,有川正俊:日本の住所体系に適した分散アドレスマッチングサービス,地理情報システム学会講演論文集,Vol.9,pp.183-186,2000 年 9 月.

(2004.5.21受付)