ワールドワイドウェブを利用した住所探索

佐藤理史[†,††]

与えられた名称から，その名称に対する住所情報をワールドワイドウェブを利用して探し出す方法について述べる．まず，検索エンジンを利用して，住所情報が記載されている可能性が高いウェブページを収集する．次に，それぞれのページに対して簡単なレイアウト解析を適用して，住所情報が記述されている領域を切り出し，そこから住所データ（名称，住所，郵便番号，電話番号，URL，出典）を抽出する．最後に，こうして得られた多数の住所データを整理・統合して，調査対象の住所情報を出力する．この整理・統合を行う方法として，属性の識別能力に基づく同一性判定という新しい方法を考案した．住所探索に要する時間は，典型的には1分から2分であり，ウェブ上に住所情報が多数記載されているものならば，多くの場合，正しい住所情報を出力することができる．

Address Search on the World Wide Web

Satoshi Sato[†,††]

This paper proposes a method that finds the address information of the given name from the World Wide Web. The method consists of five steps. The first step collects URLs of the candidate pages that maybe contain the address information of the target by using search engines. The second step downloads the candidate pages. From each page, the third step extracts the regions that contain the address information by using simple layout analysis. From each extracted region, the fourth step extracts a set of address information as a record that consists of six fields (name, address, zip code, phone number, URL, and source page's URL). The final step, the most important step, integrates many extracted records into a few reliable results by using a new method that determines equality of two records. The current implemented system can find addresses of wide variety of targets, such as company, hotel, and restaurant. The typical search time is one to two minutes.

1. はじめに

我々は日常生活において，しばしば，ある名称に対する住所や電話番号といった情報（住所情報）を調べる必要が生じる．このような場合，我々は住所録や電話帳などのリファレンス・ブックを用いて求める住所情報を探すのが普通である．これらのリファレンス・ブックは電子化された形で提供されることも多く，たとえばウェブ上では，インターネットタウンページ☆が利用可能である．リファレンス・ブック（あるいは，その基となるデータベース）は，通常多大な労力をともなう編集作業によって作成される．この編集作業のおかげで，我々は比較的容易に，求める情報を得ることができるのである．

一方，このような方法で求める住所情報を入手できない場合，ウェブの検索エンジンを用いて，ウェブページに記載された住所情報を探すという方法がある．この方法は，リファレンス・ブックには掲載されていない情報や最新の情報が得られるという利点がある一方，検索エンジンで見つかったページを1つずつ丹念に調べて住所情報を見つけ出さなければならないため，効率が非常に悪いという欠点を持つ．

ここに1つ，挑戦に値する課題を見つけることができる．

Q ウェブページの情報を利用することで，リファレンス・ブック（データベース）なしで，名称から住所情報を自動的に探し出すシステムを実現できないか．

これは，次のように言い直すこともできる．

Q' ウェブを，住所情報の仮想的なリファレンス・ブック（データベース）とすることはできないか．

† 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

☆ http://ito.ne.jp/townpage.shtml

本論文では，実際に動作するシステムとその実現法を示すことにより，課題Qに対する答えを与える．ここで用いる方法は，次の3つの技術を組み合わせたものである．

(1) 情報源の発見：住所情報が記載されているウェブページを見つける．
(2) 情報抽出：そのページから住所情報を抽出する．すなわち，ページに記載された住所情報を定型データに変換する．
(3) 情報統合：複数の情報源から見つかった住所情報を整理・統合する．

ここで一番難しいのは，(3)の情報統合である．たとえば，「佐藤病院」の住所を探索する場合を考えよう．日本全国には，たくさんの佐藤病院が存在する．情報統合では，得られたデータの中から同一の「佐藤病院」を指すデータをまとめ，いくつの異なった「佐藤病院」が見つかったのかを明らかにすることが必要となる．これは，解が一意に定まるような問題ではなく，人間が用いているヒューリスティックをうまく導入して解く必要がある．

以下，まず，2章で作成したシステムの概要について述べる．3章では情報統合について詳しく述べる．4章ではシステムの評価について述べ，5章で議論と関連研究について述べた後，6章でまとめを述べる．

2. 住所探索システムの概要

作成した住所探索システムは，与えられた名称から，その住所，電話番号，URLを探し出す．**図1**にシステムの名称入力画面を示す．**図2**に調査中の画面を，**図3**に調査結果の画面を示す．システムの応答時間は，ネットワークの状況や検索エンジンの応答時間によって大きく変化するが，典型的には1分から2分程度である．

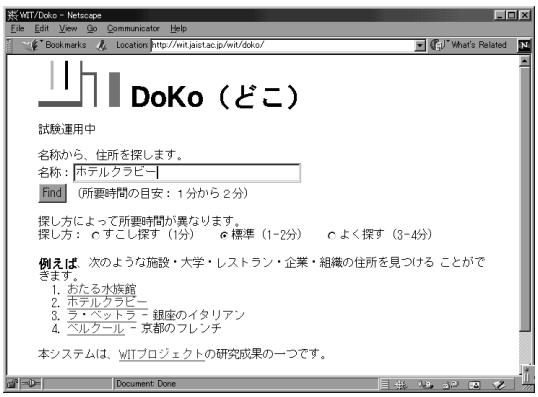

図1 名称入力画面
Fig. 1 User interface for search query.

作成したシステムの構成を**図4**に示す．本システムは，検索エンジンによる収集，ページ取得，領域抽出，情報抽出，情報統合の5つの部分から構成される．

2.1 検索エンジンによる収集

システムが最初に行うことは，与えられた名称の住所情報が記載されている可能性があるページのURLを収集することである．この収集には，既存の検索エ

図2 調査中の画面
Fig. 2 Snapshot of a search process.

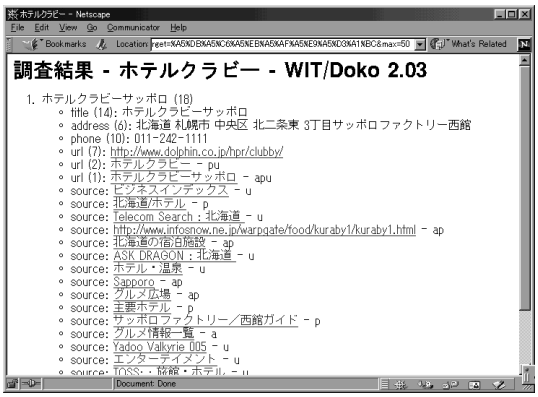

図3 調査結果
Fig. 3 Search result.

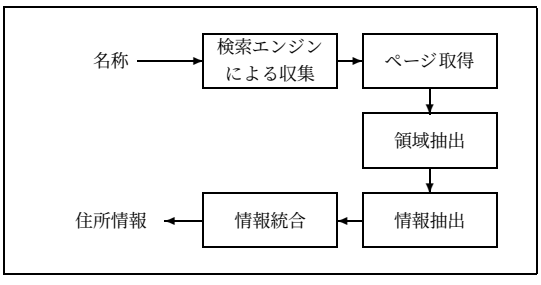

図4 住所探索システムの構成
Fig. 4 Configuration of the system.

ンジンを利用する．検索エンジンは，しばしば応答しなくなることがある．このため，複数の検索エンジンを利用する．また，各検索エンジンで，ページ収集やランキングの方法が異なるため，見つかるページの多様性確保の観点からも，複数の検索エンジンを利用することには利点がある．

与えられた名称（文字列）がウェブ上にあまり存在しない場合は，検索質問（クエリ）として「名称」を用いるのがよい．一方，与えられた名称がウェブ上に多数存在する場合は，住所を表す言葉（「住所」や「所在地」）を付加してもページが得られることが期待でき，かつ，その方が住所情報が記載されているページが見つかる可能性が高い．この両方の場合に対応するために，複数の検索質問を使用する．

現在のデフォールトの設定は以下のとおりである☆．

(1) Goo (http://www.goo.ne.jp/)
  (a) “名称”
  (b) “名称 AND (ドメインレーダー OR 所在地 OR 住所)”
(2) Infoseek Japan (http://www.infoseek.co.jp/)
  (a) “名称”

それぞれ上位 50 件の URL を収集し，最大 150 件の URL を収集する．

2.2 ページ取得

システムが次に行うことは，収集した URL のそれぞれのページを取得することである．原理的にはまったく難しいところはないが，処理時間の大部分がここで費やされるため，応答時間の短縮のためには，実装上の工夫が必要となる．

現在のシステムでは，8 プロセスによる並列取得を採用している．また，ウェブ・キャッシュ（squid 2.2）を利用している．

2.3 領域抽出

取得したページに対する処理は，領域抽出と情報抽出である．領域抽出では，ページの中から情報抽出を行う部分を特定し，その部分を切り出すことを行う．

今，あるページに調査対象の名称が記載されていたとしよう．このとき，そのページに存在する住所や電話番号は，無条件にその名称に対応するものと考えてよいだろうか．答えは否である．同じページに記載されているという条件だけから，名称と住所（あるいは，電話番号）が対応すると判定するのならば，多くの誤りを生むことになるだろう．では，どのような条件が成り立てば，対応すると判定してよいだろうか．

この問題に対する解の直観的説明は，「その名称が支配する文脈（領域）に現れるならば，その情報は，その名称に対する情報と判定してよい」というものである．今，「北陸先端科学技術大学院大学」が話題となっており，「住所は辰口町旭台 1-1」とあれば，この住所は北陸先端科学技術大学院大学の住所であると考えてよいということである．

このような文脈を把握することは一般には困難であるが，ここでは，HTML タグがある程度信頼できると仮定して，HTML タグを利用した領域把握と抽出を行う．具体的には，以下のような方法をとる．

(1) タイトルに名称が含まれる場合：ページ全体を抽出する．
(2) 見出しに名称が含まれる場合：次の同レベル以上の見出しまでの領域を抽出する．
(3) テーブルの要素やリストの要素として名称が含まれる場合：その要素を含む 1 レコード（1 アイテム）を抽出する．
(4) 文中に名称が含まれる場合：その文を抽出する☆☆．

2.4 情報抽出

こうして抽出された各領域から，住所情報を抽出する．対象領域がページ全体の場合とページの一部の場合で，方法が異なる．

対象領域がページ全体の場合は，さらに，領域の絞り込みを行う．具体的には，「住所」や「電話番号」といったラベルを探し，そのラベルの支配領域のみを切り出す．ラベルが見つからない場合は，タイトル以外の部分に現れる名称を探し，その支配領域を切り出す．

こうして切り出された領域（ページの一部の場合は，領域抽出で切り出されたそのままの領域）に対して，情報抽出を試みる．抽出する情報は，名称，住所，郵便番号，電話番号，URL，出典，の 6 種類である．

- 名称
  入力された名称が必ずしも正式名称とは限らない．略称（たとえば「上野動物園」）から正式名称（「東京都恩賜上野動物園」）が見つかるように，名称に一致した部分の前後を字種を手がかりとした規則に従って拡張し，抽出する．
- 住所
  郵政省が公開している新郵便番号のデータベースを加工して作成した住所辞書を用いて抽出する．

☆「ドメインレーダー」はドメインと組織名の対応表を提供しているサイトの名称である（http://server.nda.co.jp/）．

☆☆ より洗練された方法が望まれるが，筆者の知る限り，安定して利用できる文脈把握技術はまだない．

住所の補完機能（「石川県辰口町」→「石川県能美郡辰口町」）を持つ．

- 郵便番号
  正規表現パターンを用いて抽出する．郵便番号記号（〒）と上記の住所を文脈情報として利用し，その前後に現れるもののみを抽出する．
- 電話番号
  正規表現パターンを用いて抽出する．
- URL
  名称がアンカータグによって囲まれている場合に，そのURLを抽出する．また，名称がそのページのタイトルに含まれる場合は，そのページはその名称のURLと判断し，抽出する．後者の場合，ここで抽出されるURLと次の出典は同一となる．
- 出典
  上記の情報を抽出したページのURLを抽出する．

いずれの情報も，その領域で最初に見つかったものを採用する．こうして得られた6フィールドからなる情報を，以下では**住所データ**と呼ぶ．ただし，住所データの6つのフィールドの値はすべて抽出されるとは限らない．なお，住所，電話番号，URLのいずれも見つからなかった場合は，その領域を破棄する．

3. 情報統合

3.1 情報統合問題

前章で述べた領域抽出と情報抽出により，多数のページから多数の住所データが得られる．一例を**表1**に示す（スペースの関係上，3つのフィールドに限定した）．ここでは，7件の住所データが得られている．情報統合で行わなければならないことは，この7件のデータから，「佐藤病院はいくつ見つかったと考えるのが妥当か」ということに対する答えを決定することである．

上記の問題をここでは，情報統合問題と呼ぶことにする．この問題は，おおよそ，「独立に得られた複数のデータ（情報）から，いかにして最終的な結論を得るか」という問題である．

表1 情報抽出によって得られた住所データ
Table 1 Extracted address data.

ID	名称	住所	電話番号
#1	佐藤病院	群馬県高崎市柳川町4	0273-22-2145
#2	佐藤病院	群馬県高崎市柳川町4	0273-22-2145
#3	佐藤病院	群馬県高崎市若松町96	0273-22-2243
#4	佐藤病院		0273-22-2243
#5	佐藤病院	群馬県高崎市若松町96	
#6	佐藤病院		0287-43-0758
#7	佐藤病院	栃木県矢板市土屋18	

情報統合問題の難しさは，次の2点からもたらされる．

(1) データの不完全性：完全なデータが得られるとは限らない（欠損の存在）．
(2) データの不確実性：得られたデータがすべて正しいとは限らない（誤りの存在）．

これらの条件から，この問題は解が一意に定まるような問題ではないことは明らかである．この問題を解くということは，「得られたデータの範囲で導ける妥当な結論は何か」を求めることにほかならない．

3.2 属性の識別能力による同一性の判定

ここではまず，データに誤りのない理想的な状況を考える（ただし，データの一部に欠損は存在する）．このような状況において，表1を例にとり，情報統合問題の解法を考える．その要点は，2つのデータ（レコード）が同一の対象を表しているかどうかをいかにして判定するかということであり，より正確には，我々の持っている同一性判定に関する知識をいかにして機械に組み込むかということである．

ここで提案する方法は，属性に対して識別能力を定義し，それに基づき同一性を判定する方法である．

3.2.1 識別能力

ある属性が，対象の同一性の識別にどのように寄与するかを，次の2種類に分けて定義する．

(1) 正の識別能力：ある属性の値が一致するならば，2つのデータは同一の対象を表していると判定してよい場合，その属性は**正の識別能力**を持つ．
(2) 負の識別能力：ある属性の値が一致しないならば，2つのデータは同一の対象を表していないと判定してよい場合，その属性は**負の識別能力**を持つ．

十分条件，必要条件という言葉を用いて説明することもできる☆が，いずれにせよ，正，負という2つの方向の識別能力があることに注意されたい．

それぞれの属性は，

- 正の識別能力だけを持つ．
- 負の識別能力だけを持つ．
- 正の識別能力と負の識別能力の両方を持つ．
- どちらの識別能力も持たない．

のいずれかをとる．

3.2.2 識別能力の定義例と同一性の判定例

対象を病院として，名称，住所，電話番号に識別能

☆ ある属性の値の一致が2つのデータの同一性判定の十分条件である場合，その属性は正の識別能力を持つ．ある属性の値の一致が2つのデータの同一性判定の必要条件である場合，その属性は負の識別能力を持つ．

力を定義してみよう．

- 名称 ··· 識別能力を持たない．
  異なる病院が同一名称をとることは珍しくない．また，同一病院がいくつかの異なる名称で呼ばれることは珍しくない．
- 住所 ··· 負の識別能力を持つ．
  住所が異なれば，違う病院であると考えてよいだろう（ここでは，分院は別の対象と考える）．しかし，同一住所（たとえば同一ビル内）に複数の病院が存在することはある．
- 電話番号 ··· 正の識別能力を持つ．
  電話番号が一致していれば，同じ病院だと考えてよいだろう．1つの病院に，複数の電話番号が存在するため，電話番号が異なっているからといって違う病院と考えることはできない．

この説明から分かるように，提案手法は，人間が持っている常識を属性の識別能力という形で定義し，それを用いて同一性の判定を行うものである．

識別能力は，属性だけでなく，属性から計算できる仮想的な属性に対しても定義してもよい．ここでは，次の2つの仮想的な属性に識別能力を付与する．

- 市外局番 ··· 負の識別能力を持つ．
- 名称と住所の直積 ··· 正の識別能力を持つ．

これらの識別能力によって，先の7件の住所データ（表1）の同一性を判定すると，**表2** に示す同一性判定表が得られる☆．この表において，“+” は同一だと判定されたもの，“−” は非同一だと判定されたもの，“(+)” と “(−)” は簡単な推論☆☆を用いて同一あるいは非同一だと判定されたもの，空欄は同一性が判定できなかったものを表す．

この表から得られる帰結は，

- (#1, #2)，(#3, #4, #5)，#6，#7 の4病院か，
- (#1, #2)，(#3, #4, #5)，(#6, #7) の3病院，

のいずれか，である．これは，我々の直観に合う．

表2 同一性判定表
Table 2 Equality table.

	#1	#2	#3	#4	#5	#6	#7
#1	+	+	−	(−)	−	−	−
#2	+	+	−	(−)	−	−	−
#3	−	−	+	+	+	−	−
#4	(−)	(−)	+	+	(+)	−	(−)
#5	−	−	+	(+)	+	(−)	−
#6	−	−	−	−	(−)	+	
#7	−	−	−	(−)	−		+

☆ 判定表は右上半分で十分であるが，分かりやすさのため，冗長に示した．この表において，対角線上にある，すべてが + となっている正方形のブロックが，同一対象となる．

☆☆ 同一性の推移律と「X と Y が同一であり，Y と Z が非同一ならば，X と Z は非同一である」が成り立つ．

3.3 住所検索システムにおける情報統合

実際のシステムで用いている情報統合は，誤りを許容しなければならないため，先に述べた方法をいくつかの点で拡張した方法を用いている．

第1に，現実の問題では，属性値の同一性の判定も完全にはできない．住所を例にとろう．住所表記を完全に標準化することができれば，住所の一致，不一致を文字列の一致，不一致として判定できるが，現実には，住所表記を完全に標準化することはほとんど不可能である．電話番号は比較的同一性を判定しやすいが，大阪で3桁の局番の電話番号（06-xxx-xxxx）が見つかることもある（現在は4桁）．この問題に対処するために，属性の一致，不一致の判定を多段階で行う．

第2に，識別能力として正，負の2種類ではなく，その強さに段階を設け，強正，正，弱正，弱負，負，強負の6種類に増やす．たとえば，住所が都道府県レベルでも一致しない場合は，強負（非同一である可能性が非常に高い），住所が文字列として完全に一致する場合は弱正（同一である可能性が多少ある），のように，多段階の属性値一致度それぞれに対して，識別能力を定義する☆☆☆．最終的な同一性の判定は，これらの結果を総合して決定する．多数の識別能力を冗長に定義しておくことにより，比較的誤りに強い判定が実現できる．

第3に，相対的に頻度の低いものを棄却する．たとえば，データが10件あり，そのうち8件が同一対象を指しており，残りの2件が，それぞれ別のものを指していると判定された場合，残りの2件は雑音と考え，除去する．具体的には，以下のように行う．

(1) データ間の同一性を判定し，同一対象を表すデータをまとめる．
(2) こうして得られたグループ（同一対象を表すデータ群）をデータの件数の多い順に並べる．
(3) データ件数の減少率（となり合う2つのグループのデータ件数の比）が20%以下ならば，その間で線引きする．
(4) 先頭からのデータ件数の累計が全体の75%を超えたところで線引きする．ただし，最後のグループとデータ件数が同じものが存在する場合，それらは残す．

☆☆☆ 本来は，同一性を判定する対象の種類（病院，水族館，レストラン）に対して，それぞれ異なった識別能力を定義するのが好ましい．しかし，住所探索では，その対象が何であるかは不明であるため，デフォールト的な定義で代用する．

これにより，データ数が十分に存在する場合は，低頻度の誤りを排除することができる．なお，ステップ (3) の 20%，ステップ (4) の 75%という数字は，経験的に決定した．

4. システムの性能評価

本システムの性能をどのような方法で評価すべきかは，それほど明白ではない．現実世界に対して開かれたシステムであるため，実行結果には再現性がない（ネットワークの状況などにより同一の入力に対して異なった出力が得られる）．また，理想的な閉じた環境を作って実験することも現実的ではない．

ここでは，まず，ある入力に対するシステムの実行過程を詳細に示すことによって，システムがどのように動くのかを示す．次に，115 の入力に対するシステムの出力を評価し，システムの性能と限界を示す．

4.1 実行過程の詳細

ここでは，「文化放送」を入力した場合の，出力に至るまでの経過を詳しく示す．

(1) 2つの検索エンジンに対する 3 つの検索により，150 件の URL が見つかった（重複なし）．このうち，137 ページが取得できた．

(2) これらのページに対して領域抽出を行い，330 件の領域を抽出した．

(3) 330 件の領域に対して情報抽出を行い，96 件の住所データを抽出した．

(4) 96 件のデータの同一性を判定したところ，45 のグループが得られた（同一だと判定できなかった場合は，非同一だと判定する）．このうち，データ件数が 1 であるグループを除いた 5 グループを**表 3** に示す．この表の右端の数字はそのグループに属する住所データの件数，または，その値を支持する住所データの件数を表す．

(5) 4 位のグループは，URL しか得られていないので排除される☆．4 位が削除された後，3 位のグループと 5 位のグループのデータ件数の減少率を求めると 2/12 = 17%となり，20%を下回るので，最終的には，上位 3 位のみが出力される．

こうして，文化放送（東京），北海道文化放送，長崎文化放送，の 3 つの住所情報が探索結果として得られた．

4.2 システムの評価

本システムに 115 の名称を入力し，その結果を整理したものを**表 4** に示す．ここでは，出力の評価として，以下の 5 段階の評定を用いた．

A（優） 単独 1 位☆☆から，調査対象の住所，電話番号，URL のすべてが得られる．

B（良） 1 位から，調査対象の住所，電話番号，URL のうち，2 つ以上が得られる．

C（可） 1 位から，調査対象の住所，電話番号，URL のいずれかが得られ，かつ，それ以外の要素に明らかな誤りが存在しない．

F 出力が得られたが，A〜C に該当しない．

N 何も出力が得られない．

なお，この表の名称の後の数字は，出力の 1 位のデータ件数（その対象を表すと判定された住所データの件数）を表す．また，'*' は，検索エンジンに対する 3 つの検索質問のうち 1 つが，答を返さなかった（検索で

表 3 「文化放送」の調査結果（上位 5 位まで）
Table 3 Search result for "Bunka-Housou" (top five).

1	文化放送		21
	名称	文化放送	14
	郵便番号	160-8002	2
	住所	東京都新宿区若葉 1-5	2
	電話	03-3357-1111	2
	URL	http://www.joqr.co.jp	21
2	北海道文化放送		15
	名称	北海道文化放送	8
		北海道文化放送 (UHB)	3
		北海道文化放送株式会社	2
	郵便番号	060-8527	1
		060-0001	1
	住所	北海道札幌市中央区北一条西 14 丁目	2
	電話	011-214-5200	2
	URL	http://www.uhb.co.jp/	15
3	長崎文化放送		12
	名称	長崎文化放送	10
	郵便番号	852-8527	2
	住所	長崎県 長崎市 茂里町 3-2	2
	電話番号	095-843-1000	2
	URL	http://www.ncctv.co.jp/	12
4	長崎文化放送		6
	名称	長崎文化放送	3
		長崎文化放送 (NCC)	2
	URL	http://www.tv-asahi.co.jp/network/NCC/index-j.html	6
5	文化放送ブレーン		2
	名称	文化放送ブレーン	1
		文化放送ブレーン営業局	1
	電話番号	03-3578-3111	1
		03-3578-3161	1

☆ 最終的に残されるものは，住所あるいは電話番号が存在する場合，および，URL がホスト名となっている場合のみである．

☆☆ ただし，同名の別対象が存在する場合は，必ずしも 1 位である必要はないとする（B，C も同様）．この例外に該当したのは，今回の実験では，「文化放送」と「あさひ幼稚園」のみ．前節の実行過程の調査では，「文化放送」が 1 位となったが，この実験の場合は，「北海道文化放送」が 1 位となった．

表 4 システムの評価
Table 4 Evaluation.

A – 32 件
北陸先端科学技術大学院大学 (77), 上野動物園 (44), 金沢学院大学 (42), 札幌大学 (42), 日本点字図書館 (42), 国立西洋美術館 (40), 石川県庁 (37), サントリー美術館 (36), 石川県立図書館 (36), 円山動物園 (34), おたる水族館 (33), 北沢書店 (25), クアハウス九谷 (25), 郵政省 (24)*, 北陸経済研究所 (22), 科学技術庁 (19), 文化放送 (18), ホテル日航金沢 (14), 鶴来町役場 (13), 文部省 (13), オホーツク水族館 (12)*, たばこと塩の博物館 (12), 紀ノ国屋 (11), ラ・ベットラ (11), 東京ドーム (10), クラビーサッポロ (7), 竹香 (7), 鹿島建設 (6), 夢の島熱帯植物園 (5), つる幸 (3)*, カストール (3), 鎌倉文学館 (3)*
B – 30 件
ジュンク堂書店 (30), 日本銀行 (23), 小松税務署 (21), 中原中也記念館 (16), 松戸保健所 (12)*, 金沢全日空ホテル (12), 金沢大学医学部附属病院 (12), バードハミング鳥越 (10), 石川厚生年金会館 (9), 松戸市役所 (9)*, 室蘭水族館 (9)**, 金沢赤十字病院 (8), オーチャードホール (8), ソニー (8)*, 早稲田松竹 (8), 角川書店 (7)*, 青山円形劇場 (6)*, ギャラリー青雲 (4)*, 四川飯店 (4), 電力中央研究所 (3), 博報堂 (2), ACB ホール (2)*, 川村記念美術館 (2)*, 石川テレビ (2)*, 高見小学校（北九州市）(1)*, 県立船橋高校 (1), JTB 金沢支店 (1)**, 松戸北郵便局 (1)*, 久保書店 (1), いしかわ動物園 (1)**
C – 25 件
北国新聞 (36), 歴史民族博物館 (34), 朝日新聞 (20)*, 京都大学 (14), 新宿スカラ座 (10)*, 電通総研 (8), ナムコワンダーエッグ (8), 岩手県立大学 (6)**, 辰口町役場 (6), 岩波書店 (6)*, 最高裁判所 (6)*, フジテレビ (4)*, 富士急ハイランド (4)*, 草月ホール (3), あさひ幼稚園（松戸市）(3), 中山競馬場 (2), 蘭東中学校（室蘭市）(2), オーム社 (2)*, 国税庁 (2)**, サイゼリヤ (2)*, 春光小学校（旭川市）(1), 室蘭栄高校 (1), 日動火災 (1)*, 総合ビジョン (1)**, 四川一貫 (1)*
F – 23 件
北海道庁 (23), NHK (14)*, 高島屋 (13)*, ナナオ (12), 北海道大学 (9)*, 講談社 (6), 新日本製鉄 (5), 広島ビッグアーチ (5)*, 銀座松屋 (4)*, ベルクール (4)*, BMW ジャパン (4)*, 鳳舞 (3), 八条中学校（札幌市）(3), 丸紅 (3)*, マキシム・ド・パリ (3)*, 小松空港 (2)*, 河原塚中学校（松戸市）(2)*, 船橋競馬場 (2), 寺井警察署 (2), 帝国劇場 (2)*, 北国銀行 (1)*, 金沢駅 (1)*, 辰口交番 (1)
N – 5 件
札幌南高校*, 知利別小学校（室蘭市）**, 辰口町立図書館**, つくば第一ホテル, 銀座はげ天

きなかった）ことを表し，‘**’は，3 つのうち 2 つが答を返さなかったことを表す．

評点 A，B，C は，それぞれ，優，良，可に相当するものとして設定した．A および B の場合は，ウェブを仮想的なリファレンス・ブックとすることに成功していると考えてよい．C は，少なくとも通常の検索エンジンよりは有効に働くことを意味する．

これらの結果から，おおよそ以下のような傾向があることが分かる．

- ウェブから多数の住所データを得ることができた場合に，正しい住所情報が得られることが多い．そもそもウェブ上に住所情報の記述が少ない場合（小中学校）や，検索エンジンによる検索に失敗した場合に，正しい住所情報が得られない可能性が高くなる．たとえば，「辰口町立図書館」は，3 つの検索質問のすべてに対して答が得られた場合には，評定 A の出力を返す．
- 本システムが得意とするのは，動物園，水族館，美術館などの施設である．これは，それらの住所一覧がウェブ上に比較的多数あるためと考えられる．
- 本システムがほとんど有効に働かないのは，駅や空港である．たとえば，「金沢駅徒歩 5 分」のような表現の中に「金沢駅」を見つけてしまい，誤った住所を抽出する．また，駅や空港は，そもそも住所情報がそれほど重要な情報ではないため，ウェブ上にあまり記載されていないことにも有効に働かない原因の 1 つである．
- 大きな企業や大学，多くの支店を持つものなどは，比較的苦手である．これは，複数の住所や多数の電話番号を持つためである．
- 略称や別表記をある程度許容する（たとえば，評定 A の「おたる水族館」は「小樽水族館」でも評定 B の結果が得られる）が，正式名称を入力した方が結果は良い．たとえば，評定 F の「NHK」，評定 N の「つくば第一ホテル」は，それぞれ「日本放送協会」，「筑波第一ホテル」ならば，評定 A の出力を返す．
- 情報統合は多くの場合有効に機能するが，しばしば誤った同一化を行ったり，正しいデータを雑音と判定して棄却したりすることもある．たとえば，「鳳舞（京都市）」の住所データは発見されたが，他に頻度が高い別対象が見つかったため，棄却された．

5. 議論と関連研究

システムの性能評価（4 章）で示したように，本システムはいろいろなカテゴリの対象の住所情報を見つけることができる．このことから，1 章で提示した課題 Q は達成されたと考える．

本システムが行っている処理は，ウェブを仮想的なリファレンス・ブック（仮想的なデータベース）とする技術ととらえるのがよい．この技術は，生のデータからリファレンス・ブックを編集する作業のうち，データを収集する，データを整理する，データの信頼性を高める，などかなりの部分をカバーする．本研究の延長線上に，ウェブから住所録などのリファレンス・ブックを自動生成することが考えられる．しかし，このためには，対象を網羅的に集めることや，どれを載せどれ

を載せないかを決定する，などの本システムでは扱っていない処理が必要となる．

本研究は，知的ソフトウェア，データベース，情報抽出などの研究と関連があるが，最も関連が深いのは，ウェブ上の知的ソフトウェア（情報エージェント）の研究である．本システムは，Etzioni のいうところの information carnivores [1]に分類できる．あるいは，検索する情報の種類を限定したメタ検索エンジン[2]という見方もできる．複数の情報源からの情報を利用する知的ソフトウェアでは，それらの情報をまとめるために，何らかの形で情報統合を行う必要がある．特に，ウェブの世界では，大量かつ冗長に情報が存在する一方，誤った情報も多い．本システムに組み込んだ情報統合は，この問題に対する 1 つの解を与えている．

情報統合という言葉は，現在，いろいろな意味で使われている（たとえば，文献 3)，4)）．複数のデータが表現している対象（実体）の同一性を判定し，それらをマージするという意味での情報統合を自動化したものに，伊藤らの事例に基づくフレームマッピング[5]がある．この方法は，異なるフレーム間のマッピング（より正確には，同じ属性を表すのにもかかわらず，スロット名が異なる 2 つのスロットを対応付けること）に主眼が置かれており，それを行う際に副次的に 2 つのインスタンス（データ）の同一性が判定される枠組みとなっている．彼らが扱っているスロットの対応付けの問題は，本研究では情報抽出によって吸収されているため，情報統合においては発生しない．また，彼らは同一性判定に汎用的な尺度を用いているが，本研究では，人間の持つ領域知識を利用する立場をとっている．これらの点に大きな違いがある．

一方，データベースや情報抽出の分野では，情報統合は重要な問題として認識されてこなかった．データベースの分野では，適切にモデル化された理想的な世界が扱われ，そこでは，対象（実体）とそれを表しているデータの同一性が保証される．そのため，本研究で扱ったような情報統合の問題は現れてこない．情報抽出（たとえば，文献 6)）では，テキストに書かれている情報を抽出することに力点が置かれ，抽出した情報を整理することには，いまのところほとんど関心が払われていない．また，テキストに誤った情報が記述されているといった可能性はまったく想定されていない．このため，情報統合の問題は扱われてこなかった．

6. ま と め

本論文では，ある名称から，その名称に対する住所情報をウェブを利用して自動的に見つけ出すシステムについて述べた．現在のシステムは，応答速度に関して若干問題を残しているが，ほぼ実用レベルに達している．本システムの前身は，リンク集自動生成システム[7]の一部であり，すでに数千件以上の住所情報を発見した実績を持つ．また，本システムの最初の版が動き出してから，すでに半年以上経過しているが，その間，実際に住所を調べるために何度も役立っている．

ウェブを仮想的なリファレンス・ブックとするための最大の問題は情報統合であった．属性の識別能力に基づく同一性判定法は，この問題に対する解を与えた．この方法は，我々人間が持っている同一性判定のための知識を，属性の識別能力という形で容易に組み込むことができ，こうして組み込まれた知識に基づきデータ間の同一性を判定することによって，ウェブから抽出した多数のデータを整理することができる．この方法は，ウェブから抽出したデータ以外にも広く適用することができるため，たとえば，既存のデータベースの検索結果として得られたデータと，ウェブから抽出したデータの両者を考慮して最終結果を決定することは容易に実現可能である☆．

謝辞 システムのテストと評価を手伝ってくれた宇夫彩子さんに感謝します．なお，本研究は，筆者が北陸先端科学技術大学院大学に所属していたときに行ったものである．

参 考 文 献

1) Etzioni, O.: Moving Up the Information Food Chain, *AI Magazine*, Vol.18, No.2, pp.11–18 (1997).
2) Sleberg, E. and Etzioni, O.: The MetaCrawler Architecture for Resource Aggregation on the Web, *IEEE Expert*, Vol.12, No.1, pp.11–14 (1997).
3) 松原 仁, 岡 隆一（編）：小特集：「情報統合への視点」, 人工知能学会誌, Vol.11, No.2, pp.176–208 (1996).
4) 竹澤寿幸（編）：論文特集：「マルチモーダル情報統合システム」, 人工知能学会誌, Vol.13, No.2, pp.205–251 (1998).
5) 伊藤史朗, 上田隆也, 池田裕治：分散情報源に対する情報エージェントのための事例に基づくフレームマッピング, 電子情報通信学会論文誌（D–I), Vol.J81-D-I, No.5, pp.433–442 (1998).
6) Defense Advanced Research Projects Agency: *Proc. 6th Message Understanding Conference* (*MUC-6*), Morgan Kaufmann Publishers (1995).

☆ システムには，その機能はすでに組み込まれているが，システムの評価は，既存のデータベースを利用せずに行った．

7) Sato, S. and Sato, M.: Toward Automatic Generation of Web Directories, *Proc. International Symposium on Digital Libraries 1999* (*ISDL '99*), pp.127–134 (1999).

(平成 12 年 1 月 31 日受付)
(平成 12 年 11 月 2 日採録)

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 年から 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），「情報の組織化」（共著，岩波書店，2000）等．