ワールドワイドウェブを利用した住所探索

佐　藤　理　史

北陸先端科学技術大学院大学／さきがけ研究21

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・ブックを用いて調べるのが普通である。一方、このような方法で見つからない場合には、ワールドワード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をとることもできる。この方法は、リファレンス・ブックに掲載されていない情報や最新の情報を得ることができるという長所がある一方、検索エンジンで見つかったページを一つづ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短所を持つ。

本論文では、この効率の悪い作業を自動化する方法、すなわち、与えられた名称から、その名称に対する住所情報をウェブから自動的に探し出す方法を示す。これは、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）とすることに相当する。

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。

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、ダウンロード、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なうことは、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。

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。現在は、2つの検索エンジンに対して合計3回の検索を行ない、それぞれ上位50件（最大150件）のURLを収集する。

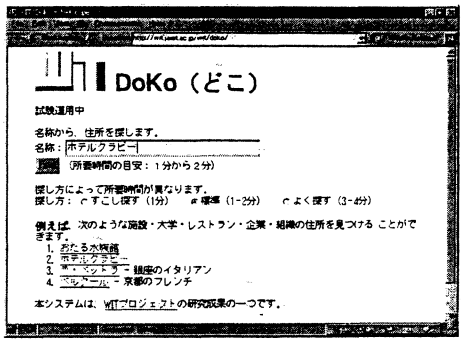

図1　名称入力画面

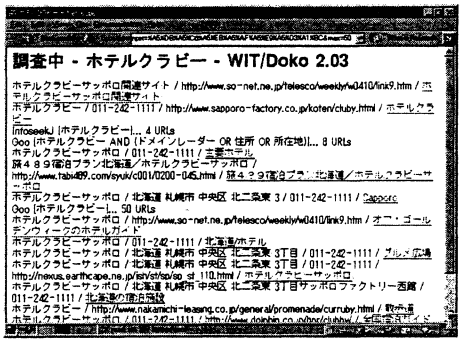

図2　調査中の画面

2.2 ダウンロード

システムが次に行なうことは、収集したURLのそれぞれのページのソースをダウンロードすることである。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ダウンロードを採用している。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

ダウンロード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、ページの中に存在する

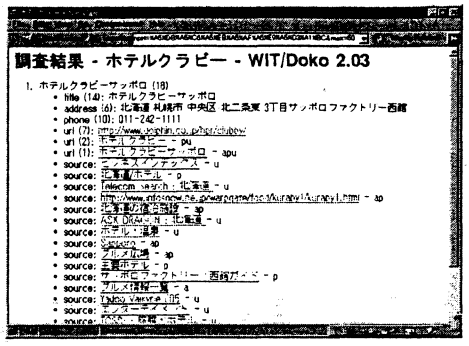

図 3　調査結果

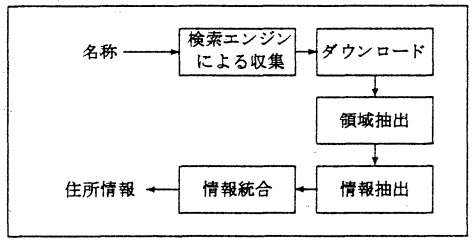

図 4　住所探索システムの構成

名称を見つけ、それが支配する領域を HTML タグを手がかりにして抽出する。具体的には、以下の方法を用いる。

(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には、情報抽出に先立ち、さらなる領域の絞り込みを行なう。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（例えば「上野動物園」）から正式名称（例えば「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。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。
- 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。但し、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。

表 1　情報抽出によって得られた住所データ

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	

3. 情報統合

前節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。情報統合で行な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。

この問題の難しさは、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」を求めることに他ならない。

3.1 属性の識別能力による同一性の判定

ここでは、対象の同一性の識別のために、属性に以下の2種類の識別能力を設定する。

(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。

それぞれの属性は、

- 正の識別能力を持つ
- 負の識別能力を持つ
- 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。

例えば、対象を病院として、名称、住所、電話番号に識別能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。一つの病院に、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。

この説明からわかるように、ここで提案する方法は、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なうものである。

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。例えば、

- 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、

表2 同一性判定表

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

- (#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。

3.2 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、先に述べた方法を以下の3点で拡張した方法を用いている。

(1) 属性値の一致、不一致の判定を多段階で行なう。これは、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ことに対処するためである。

(2) 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。例えば、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。

(3) 相対的に頻度の低いものを棄却する。例えば、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。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。

4. 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3**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。

---

* 「XとYが同一であり、YとZが同一ならば、XとZは同一である」などが成り立つ。

** 但し、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。

表 3 システムの評価

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

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 位のデータ数を表わす。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良い。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。

5. まとめと関連研究

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から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 [1] 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。

本研究は、知的ソフトウエア、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エア（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 [2] 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 [3] 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エアでは、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なう必要がある。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、この問題に対する一つの解を与えている。

情報統合という言葉は、現在、色々な意味で使われている。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 [4] 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、それを行なう際に副次的に 2 つのインスタンス（データ）の同一性が判定される枠組となっている。彼らが扱っているスロットの対応づけの問題は、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。

参 考 文 献

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

2) Etzioni, O.: Moving Up the Information Food Chain, *AI Magazine*, Vol. 18, No. 2, pp. 11–18 (1997).

3) Sleberg, E. and Etzioni, O.: The MetaCrawler Architecture for Resource Aggregation on the Web, *IEEE Expert*, Vol. 12, No. 1, pp. 11–14 (1997).

4) 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).