店舗に関する諸情報の Web からの自動収集整理提供システムに関する検討

中原　史貴　　梅原　直樹　　加藤　誠巳
（上智大学理工学部）

１．まえがき

カーナビを始めとしたナビゲーションシステム技術の進歩は目覚しく、特に店舗に関する諸情報のデータベースは増えていく一方である。しかしそのデータベースを作成するにあたっては、膨大な労力を必要とする。また、個々のユーザが必要とする情報は偏っていることが多いので、たとえ膨大な情報があったとしても、全体のごく一部の情報を扱うにすぎない。本稿では Web をデータベースとしてユーザの所望する店舗の諸情報を自動取得し、ナビゲーションシステムに応用することに関し検討した結果について述べる。

２．システムの目的

本システムでは、歩行者ナビゲーションで利用することを念頭において店舗の住所、郵便番号、電話番号、定休日、営業時間等の情報に関し、Web を用いることにより自動的に抽出、整理することで最も適した結果をユーザに返すことを目的としている。この際、ユーザがストレスを感じない時間内に結果を返すことが重要となる。

３．システムの概要

システムの大まかな流れを図１に示す。

3.1. ユーザによる店舗名の入力

始めにナビゲーションシステムを利用している状態でユーザが「店舗名」を入力する。「店舗名」が大手チェーン店などの場合には、サブキーワードとして「支店名」についても入力する。また、アプリケーション側が「大まかな現在地(区市町村)」を割り出す。

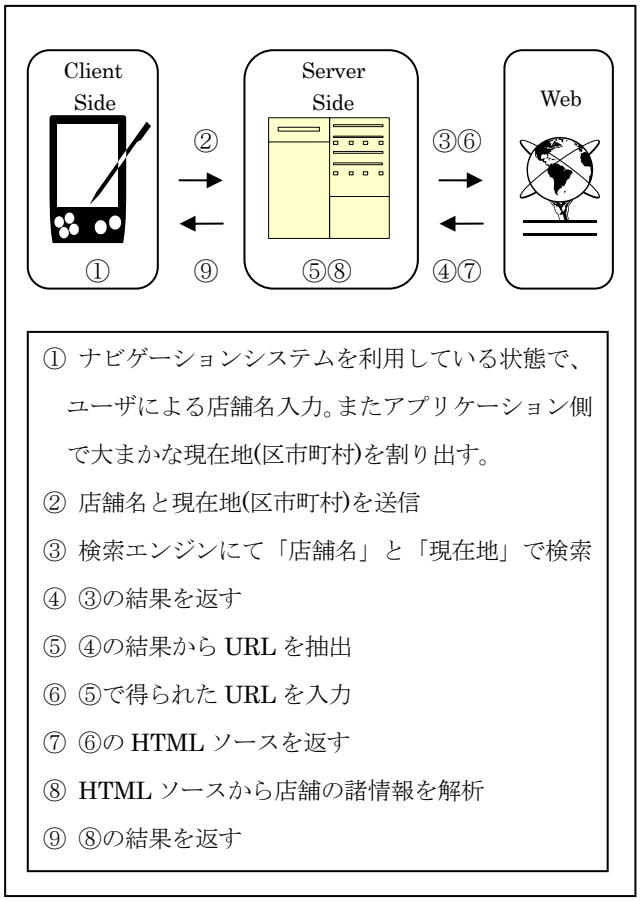

図１　システムの概要

3.2. Web からのページ取得

前項により得られたキーワードと現在地を元に店舗の諸情報が記載されている可能性のある Web ページの URL を取得する。その URL の収集には既存の検索エンジン Google を用いる。この際の検索キーワードとして「大まかな現在地(区市町村)」および「店舗名」とする。「支店名」は検索キーワードとして使うのではなく後述の情報抽出の際に用いる。この検索結果として上位 10 件を採用することにす

A Shop Information Collection System Using WWW
Fumitaka NAKAHARA, Naoki UMEHARA,
Masami KATO
Sophia University

る。これらの URL をもとに HTML ソースを取得する。ここで 10 件と少なくした理由は、このシステムの中で HTML ソースの取得に最も時間を要してしまうことと、「(現在地)AND(店舗名)」という検索キーワードから上位 10 件で精度の高い結果が得られることが多いためである。

3.3. 情報の抽出法

前項により得られた HTML ソースから店舗の諸情報を抽出する。その際、図２に示す基準に従って重要度を定める。

①タイトルに店舗名が含まれている：重要度　大
②Web 上で表として現れる箇所に店舗名が含まれている：重要度　中
③その他の箇所で店舗名がある：重要度　小

図２　情報の重要度

これらの重要度を考慮した上で図３のようにソース内の店舗名(支店名が入力されている場合は支店名)について検索し、その店舗名(支店名)から次の店舗名(支店名)が現れる間までの領域で最初に現れる対象となる情報(例えば住所)のみを抽出する。

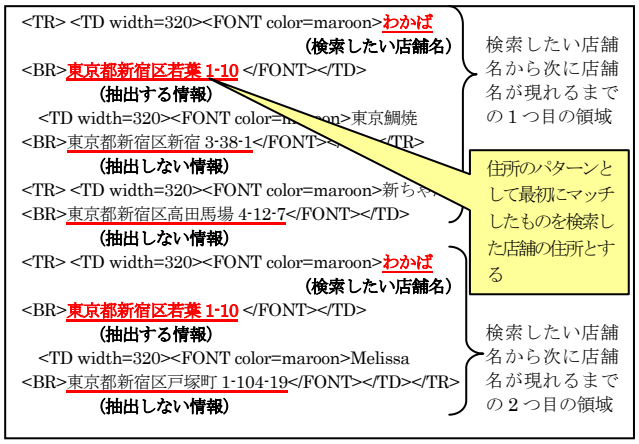

図３　対象となる情報(住所)の抽出法

3.4. Jakarta ORO [1]

上記のような情報を抽出するにあたり Jakarta プロジェクトによる Jakarta ORO を利用した。Jakarta プロジェクトは Apache プロジェクトのサブプロジェクトで、The Jakarta ORO Java クラスは、Perl5 互換の正規表現による置換、分割、ファイル名のフィルタリング等を行うユーティリティを提供する、テキスト処理に関するクラスのセットである。

3.5 情報の整理

得られた情報から重要度と抽出した件数により、その諸情報の確からしさにランキング付けを行い、最上位の情報のみ(絞りきれなかった場合には上位数件)を結果として返し、ナビゲーションシステムにおいて地図上で表示する。

４．実行例

四ツ谷駅においてナビゲーションシステムより「わかば」という店の検索をした場合について、情報収集画面を図４に示す。

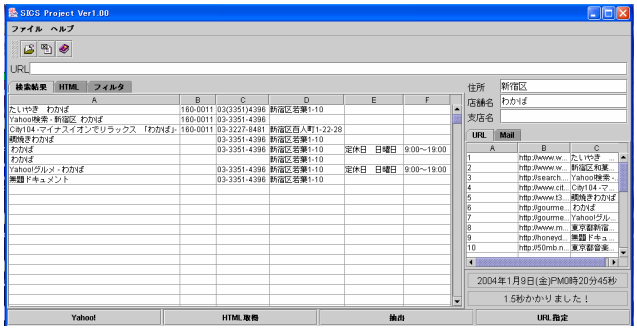

図４　情報収集画面

５．むすび

今回 Web をデータベースとして HTML で記述されたソースから店舗に関する諸情報を抽出したが、HTML は曖昧な構造を持つ言語であり、その抽出法は自ずと Web ページのレイアウトから推測したものとなる。今後、HTML に代わり厳格な入れ子の構造を成す XML への以降が進むようになれば、このシステムはより正確かつ迅速に動作し、実用に耐え得るものになると考えられる。

最後に、有益な御検討を戴いた本学 e-LAB/マルチメディア・ラボの諸氏に謝意を表する。

参考文献

[1] http://jakarta.apache.org/oro/