7V-3

自動生成した Wrapper を用いた Web 情報の自動収集・提供システムに関する検討

増田　雄紀　　加藤　誠巳
（上智大学理工学部）

1. まえがき

現在、Web 上には HTML を用いて記述された膨大な量の情報が存在しており、PC やモバイル端末を利用することで、多くの有用な情報を得ることができる。それに伴い、ユーザの代わりに膨大な量の情報を効率よく処理し、個人のニーズに特化した情報を収集し提供するサービスが求められている。

本稿ではニュースサイト、鉄道・天気サイト、Blog から情報の収集を行い、メール、Web を用いて情報提供するシステムに関し検討した結果について述べる。膨大な量の HTML から情報を抽出するためには、効率的な抽出手法が必要とされる。ニュースサイトなどに代表される特定のテンプレートを用いたサイト、鉄道・天気サイトに代表されるテーブルタグを用いたサイトに対して自動で Wrapper を生成するアルゴリズムを提案するとともに、ニュースサイト、Blog から抽出した情報を解析することで、ユーザのニーズに特化した情報の提供を行う情報システムの構築を行った。

2. システム概要

2.1 システムの構成

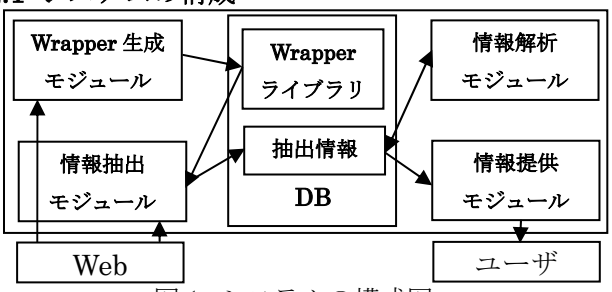

図１　システムの構成図

図１にシステムの構成図を示す。本システムは Wrapper 生成、情報抽出、情報解析、情報提供の４つのモジュールならびにデータベース（MS SQL Server）によって構成されている。

- Wrapper 生成モジュール：コンテンツ抽出 Wrapper、テーブル抽出 Wrapper を生成する
- 情報抽出モジュール：HTML から Wrapper を用いて情報を抽出する
- 情報解析モジュール：形態素解析ツールの「茶筌」を用いて抽出情報の形態素解析を行い、情報ごとに名詞を検出する
- 情報提供モジュール：ニュース、Blog は Web、鉄道、天気はメールを用いて情報を提供する

2.2 情報抽出の流れ

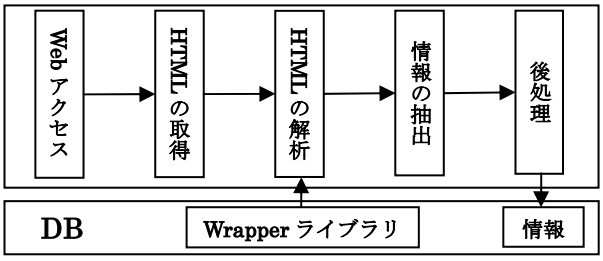

図２　情報抽出の流れ

図２に情報抽出の流れを示す。情報抽出モジュールでは URL と Wrapper を入力することで情報を抽出することができる。

- Web アクセス：URL にアクセスする
- HTML の取得：HTML を取得する
- HTML の解析:HTML パーサを用いて解析する、ライブラリから Wrapper を入力する
- 情報の抽出：Wrapper を用いて情報を抽出する
- 後処理：不要なタグや空白を除去する

HTML パーサは HTML を整形式へと整えるツールである。本システムでは NekoHTML を用いて HTML の一般的なタグ構造の誤りなどを修正し、解析を行っている。

2.3 提供情報

NIKKEI NET を特定のテンプレートを用いたサイト例、Yahoo!天気情報をテーブルタグを用いたサイト例とし、Wrapper 生成、情報提供を行う。

3. Wrapper 生成アルゴリズム

- Wrapper : HTML から情報を抽出するためのプログラム、または抽出する場所を指示する文法
- テンプレート部分とコンテンツ部分：NIKKEI NET のサイトを例にとると、カテゴリ、広告などがテンプレート部分、記事がコンテンツ部分である

3.1 LCS アルゴリズム[1]

LCS アルゴリズムとは２つの文字列の最長共通部分 LCS（Longest Common Subsequence）を求

An Automatic Information Collection and Distribution System Using Automatically Generated Wrapper
Yuki MASUDA, Masami KATO
Sophia University

めるためのアルゴリズムであり、文字列 S、T に対して S の i 文字目までの接頭文字列を S_i、T の j 文字目までの接頭文字列を T_j とするとき、$LCS(S_i,T_j)$ を部分最適解とした場合 $LCS(S_i,T_{j-1})$、$LCS(S_{i-1},T_j)$、$LCS(S_{i-1},T_{j-1})$のいずれかを再帰的に使用することによって求まる。

3.2 コンテンツ抽出 Wrapper

HTML の中からコンテンツ部分を抽出するための Wrapper である。Wrapper 生成の流れを以下に示す。図３に生成した Wrapper を示す。

1. 同種のテンプレートを用いた HTML を複数入力する
2. 入力した HTML を２つ取り出す
3. NekoHTML を用いて HTML を解析する
4. HTML から要素名、コメント、テキストを取り出し、LCS アルゴリズムを用いて入力した２つの HTML の最長共通部分列を求める
5. 求まった最長共通部分列からテンプレートとコンテンツの左・右区切れ部分を検出する
6. 区切れ部分は共通でない部分を２以上囲む要素名、コメント、テキストである
7. ２～６を全ての入力に対して行う
8. 抽出した左・右区切れ部分のペアの中で頻度の高いものを Wrapper とする

```
左：<!--FJZONE START NAME="MIDASHI" -->
右：<hr class="vis2">
```

図３ コンテンツ抽出 Wrapper

3.3 テーブル抽出 Wrapper[2]

HTML の中からテーブル構造を抽出するための Wrapper である。Wrapper 生成の流れを以下に示す。図４に生成した Wrapper を示す。

1. HTML を入力する
2. NekoHTML を用いて HTML を解析する
3. HTML 中の<table>、<tr>、<td>タグを抽出する（複数ある場合は全て抽出する）
4. <table>タグまでのパスを Wrapper とする

```
HTML-BODY-CENTER-P-TABLE-TR-TD-
TABLE-TR-TD-TABLE-TR-TD-TABLE
```

図４ テーブル抽出 Wrapper

4. 情報解析

形態素解析で検出した名詞の中で出現頻度の高いものはキーワードとして抽出情報と一緒にデータベースに登録する。また、ユーザが抽出情報を見ると同時にそのキーワードをユーザ情報に登録し、ニュース、Blog の嗜好判断に用いる。

5. 実行例

ユーザ登録画面でメールアドレス、所望する情報を登録する（図５）。登録後はユーザ画面で情報を管理することができ（図６）、リンク先をクリックすることで詳細な情報を見ることができる（図７）。ユーザが嗜好すると思われる情報はユーザ画面で上位に表示し、またメールで送信されて携帯電話で見ることができる（図８）。

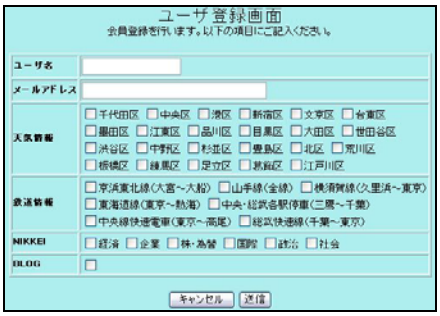

図５ ユーザ登録画面

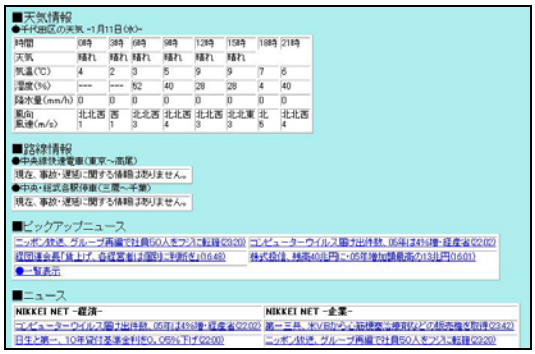

図６ ユーザ画面

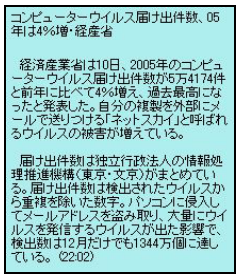

図７ 詳細画面

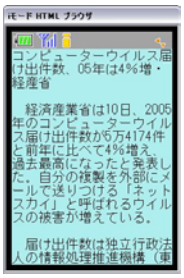

図８ メール画面

6. むすび

Wrapper を生成するアルゴリズムの提案とそれを利用した情報収集・提供システムについて検討を行った。今後は新たな Wrapper を生成するアルゴリズムを提案し、提供情報を増加させるとともに、生成した Wrapper や収集した情報を公開することで他のシステムとの連携を行いたいと考えている。

最後に、有益な御討論を戴いた本学 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]野田坂,田中:“連続する繰り返し構造を利用した Web からの情報抽出法,”情処第 66 回全大,1U-3(2004-03).

[2]増田,加藤:“個人のニーズに特化した Web 情報の自動収集提供システムに関する検討,”情処第 67 回全大,4V-9(2005-03).