5V-9

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

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、自分で Web ページを作成することにより、容易に情報を世界に発信することもできる。また、通信技術の進歩やモバイル端末の普及に伴い、場所や時間を問わず Web に接続することができ、リアルタイムで情報を得ることができる環境にある。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、鉄道、天気サイトからリアルタイムで変化する情報および Blog からユーザが望んでいる情報を自動で収集し提供するシステムについて検討を行った結果を述べる。

2. システムの概要

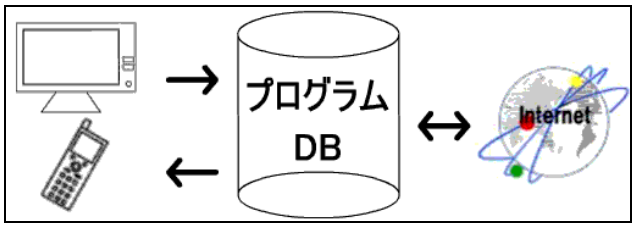

図１　システム概要図

本システムではJAVAによるプログラムでWeb上からユーザが所望する情報を抽出して提供するものである。情報の提供はメールならびに Web を用いて行っている。また、ユーザは前もってメールアドレス、所望情報などのユーザ登録を行う必要がある。抽出情報、ユーザ情報のデータ管理には SQL Server を利用している。

An Automatic Collection and Distribution System for Web Information According to Individual Preferences
Yuki MASUDA, Masami KATO
Sophia University

3. システムの詳細

3.1 システムの流れ

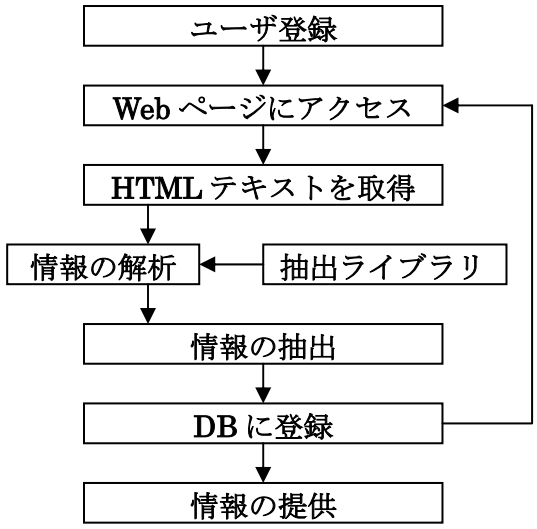

図２　情報抽出の流れ

図２にシステムの流れを示す。情報解析の部分では、抽出ライブラリを用いて収集情報の種類に合わせた方法で情報の解析・抽出を行う。

3.2 対象とした収集情報

本システムではニュース、鉄道、天気、Blog を情報収集の対象とする。

3.2.1 ニュース

Yahoo!ニュースより、カテゴリ（経済、スポーツなど）別に情報を収集する。

3.2.2 鉄道

Yahoo!路線情報より、JR の路線別に運行情報を収集する。

3.2.3 天気

気象庁の Web ページより、地域別に天気、降水確率、気温情報を収集する。

3.2.4 Blog

ユーザが望んでいる情報が掲載されている Blog を収集する。情報の収集には RSS（RDF Site Summary）を利用する。

3.3 情報の抽出

抽出ライブラリに登録されている抽出方法は、特定の Web ページの情報の抽出のみを行うのではなく、似ている HTML 構造を持つ Web ページに対して、情報の抽出ができるように作成している。

3.3.1 ニュース情報の抽出

HTML テキスト中のニュースのヘッドライン部分の繰り返し構造を検出し、記事のタイトル、掲載時刻、URL を抽出する。次にその URL から HTML テキストを取得し、タイトルと掲載時刻の間の部分を抽出する。

3.3.2 天気情報・鉄道情報の抽出

HTML テキスト中の<table>タグを検出し、テーブルタグで囲まれた部分を抽出する。<tr>タグ中に含まれる<td>タグで囲まれる要素がそれぞれ繰り返している場合は、縦に読むテーブルと考え、行と列を置換する。

3.3.3 Blog からの情報の抽出

RSS フィードから DOM（Document Object Model）を用いて、<link>、<title>、<description>タグで囲まれた部分を抽出する。<link>タグで囲まれた部分は記事の URL なので、その URL から HTML テキストを取得し、エントリー部分を抽出する。

4. 実行例

ニュース情報の抽出例を図３から図５に、鉄道情報のメール例を図６に示す。

図３ ニュースのヘッドラインの抽出

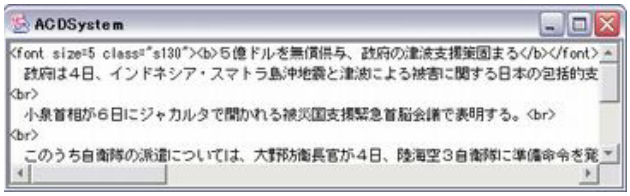

図４ ニュースの抽出

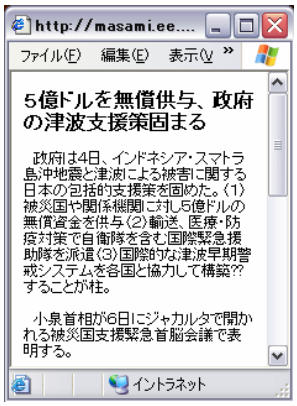

図５ ニュース出力例

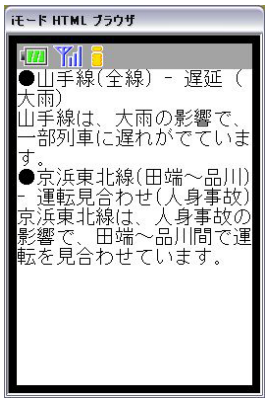

図６ 鉄道情報メール例

5. むすび

ニュース、鉄道、天気サイトからリアルタイムで変化する情報および Blog からユーザが望んでいる情報を自動で収集し提供するシステムについて検討を行った。

現段階では収集した情報をそのまま提供するだけであるが、今後は収集した情報をユーザの嗜好する情報収集にフィードバックする方法を検討し、このシステムを利用して個人のニーズに特化した情報を収集し提供することで、ユーザの日常生活やビジネスにおける情報収集に役立てたいと考えている。

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] 増田、鈴木、加藤：“Web 上の鉄道運行情報からの所望情報自動抽出提供システムに関する検討,”情処第 66 回全大,4ZB-8(2004-03).

[2] 鈴木、加藤：“所望の特定情報を Web から自動収集しメールで通知するシステムに関する検討,” 情処第 66 回全大, 4ZB-7(2004-03).

[3] 南野、鈴木、藤木、奥村：“blog の自動収集と監視,” 人工知能学会論文誌, Vol.19, No.6, pp.511-520(2004).