ウェブクリッピングによる情報収集を支援するアプリケーションの開発

越田　弘樹[I]
茨城工業高等専門学校[IV]
産業技術システムデザイン工学専攻[V]

小飼　敬[II]
茨城工業高等専門学校
電子情報工学科[VI]

滝沢　陽三[III]
茨城工業高等専門学校
電子情報工学科

1. はじめに

今日，インターネット利用者数は8500万人を超え，日本人の3分の2の人がインターネットを利用していることになる．今やインターネットは生活の一部となりつつある．そのため，人々は様々な情報と取得していくことになる．そしてその情報の整理が必要になってくる．

また，近年RSSやMicrosoft Office 2007などXMLベース形式のデータで扱うことが増え，より身近なものになってきた．

XML(eXtensible Markup Language)とは，マークアップ言語の1つで，タグを使用してテキスト形式のデータをツリー構造で表現している．そのため，XMLは簡潔で，なおかつ厳密な文法を持っているためコンピュータ上で扱いやすいという特徴を持っている．

そこで本研究は，Webから抽出した情報を簡単に整理できるようなプログラムの開発を目指す．

2. プログラムの開発

2.1概要

このプログラムは，情報の抽出，加工するプログラムと表示するプログラムの２つのプログラムで使用する．

このようにすることにより，どちらか片方のプログラムを他のプログラムに変更することができ，利便性や応用性が向上すると考えている．

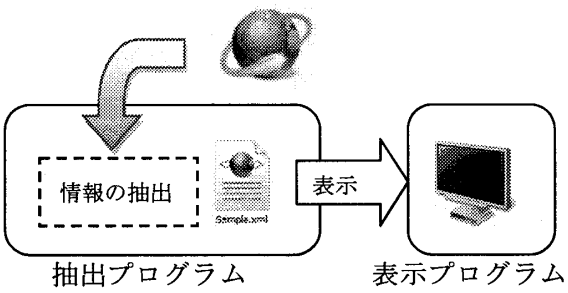

図１　プログラム構成

Fig.1 program configuration

2.2抽出・加工プログラム開発

2.2.1　データの抽出

Webページから必要な情報を抽出する．

Webページの選択した部分を，クリップボードを使用して抽出していく．クリップボードを使用する利点は，Webページのテキストデータだけでなく，ソースコードも一緒に抽出できるところにある．こうすることで，Webページのレイアウトと同じように表示することができる．[1]

2.2.2　HTMLデータをXMLデータに変換

抽出したデータはHTML形式になっている．HTMLはWebで表示するためのデータだが，XMLはデータの意味そのものを表現するためデータの整理が行いやすい．そのため，抽出したデータをXML形式に変換する．[2]

2.2.3　XMLデータの保存

変換したXMLデータを保存する．本研究のアプリケーションでは，ガジェットで表示するため，指定のフォルダに保存しなければならない．

「Development of Application to organize a HTML data by Webclipping」
I Hiroki Koshita
II Kei Kogai
III Youzou Takizawa
IV Ibaraki National College of Technology
V Advanced Course of Information Engineering
VI Electronic and Computer Engineering

2.3 表示プログラム開発

保存したXMLファイルを表示する.

今回は，表示するためのプログラムにWindows Vistaに搭載されているWindows Sidebar上で動作するガジェットを使用する. [3][4]

ガジェットとはウィジェットとも呼ばれており，他にはGoogle GadgetやYahoo! Widgets, Mac OS XのDashboardなどがある.

ガジェットで表示している利点は,普段表示している分には邪魔にならずに済むのではないかと考えている.

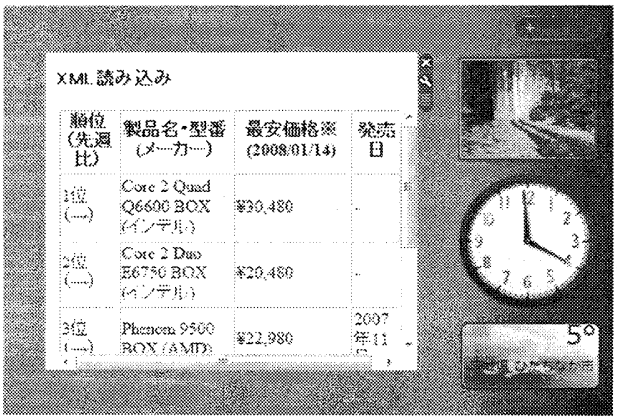

図2　ガジェット表示例

Fig.2　The indication example of the gadget

3. アプリケーションの検証

現状のWebクリッピングソフトはほとんど，テキストデータのみを抽出するものや，Webページを画像として抽出するものである.

本研究のアプリケーションが優れている面は,複数のサイトから抽出したデータを一つにまとめて表示することができるという点である.これはこのアプリケーションにおいて長所であると思われるので,これを生かしつつ機能を向上させていく予定である.

しかし,2007年10月26日にアップル社から発売されたMac OS X Leopardのなかに搭載されているSafariというWebブラウザにも，Webクリッピング機能があり,Webページをそのまま切り取り，ガジェットに張り付けることができる.さらに，画像やFlashなども切り取ることができ,本研究のアプリケーションにはない機能であり,今後の課題にしたいと思う.

4. 課題・発展

ここでは本研究のアプリケーションの課題について考察して,それに対して改良し発展させる方法を見出す.

一つは,抽出した情報の更新ができないことである.現在は一日に何度も更新することがある.よって，本研究でも自動更新やRSSのように更新状況を知らせる機能を付けることができれば,より使いやすくなると考えている.

次に，今回はWeb情報をXMLとして抽出したが,今後はテキストデータや画像の抽出だけでなく,Flashや動画など，バリエーション豊かな抽出が必要になると考えている.

あとは,ガジェットを使用しているため動作環境に依存してしまっていることである.今回はWindows Vistaに標準搭載されているガジェットを使用したが,他のガジェットとの互換性がないため,今後の研究で他のガジェットに対応するアプリケーションの開発や互換性を持ったアプリケーションの開発が必要になると考えている.

5. まとめ

本研究では，容易にWebページの情報を抽出することを目的として行ってきたで,ワンクリックで抽出から保存までできるようになっている.

今後の開発では，実際の使用による検証を行い,本アプリケーションを拡張性に優れたものにする必要があると感じた.

参考文献

[1] DotNet Programin Note.
http://fml.cside.to/program/net/ClipBordHtml_00.html
[2] CodeZine http://codezine.jp/a/article/aid/448.aspx
[3] MSDN　http://msdn.microsoft.com/
[4] ITpro
http://itpro.nikkeibp.co.jp/article/COLUMN/20070910/281556/?P=1&ST=swd-tech