Journal of National Fisheries University　55 (2) 53－58(2007)

電子メールによるWebページの取得と後処理

瓜倉　茂[1†], 佐川幸久[2], 青木邦匡[1], 楫取和明[1]

Acquiring a web page by e-mail and post-processing

Shigeru Urikura[1†], Yukihisa Sagawa[2], Kunimasa Aoki[1], and Kazuaki Kajitori[1]

This peper introduces two tools that we have developed, the first one is the main tool in order to use to acquire a web page by e-mail under the circumstance whose Internet utilization environment is harsh, and the second one is the tool to use for post-processing. If you need only the character information on the web page, you can reach the purpose with the e-mail sending and receiving of 1 times. If you need the graphics data on the same web page furthermore, e-mail operation is necessary again. Under the thought of being good that you can acquire the minimum information of the web data, the technique which is expressed with this paper is effective method.

Key words：Web, E-mail, HTML

1　はじめに

今日、インターネット接続においてADSLや光ケーブル、ケーブルTVなどを使った接続が広い地域で利用できるようになり、高速なインターネット接続サービス、つまりブロードバンドが利用できる環境が身近にあるということが当たり前になってきている[1)]。また、文字・画像を中心としたWebデータの閲覧もブロードバンドを利用した日常生活の一部となってきている。これらはIT技術の恩恵を享受できる光の部分とすれば、一方、影の部分ともいえる山間や僻地、離島などではブロードバンドの走りでもあるADSLのサービスも受けられないところもある。運航中の船舶や飛行機においても然りで、これらでは通信回線を人工衛星に頼ることになり設備費、通信費など高額な費用が必要となる場合である[2)]。

本稿では、低速なインターネット接続しか利用できない環境において、Web通信を使わないで、電子メールを通信の手段として、Webページを取得するツールを作成しその使用方法や後処理を紹介する。Webページの文字情報のみ、あるいはPDFのみ必要であるような場合であれば、1回の電子メールの送受信で十分間に合うだろう。文字以外にさらに画像が必要な場合は、再度電子メールにて画像情報を取得する必要があり、また後処理としてWebページのHTMLソースを少し手直しすることで文字・画像を同時にWebブラウザに表示させることが出来る。

2　作製したツールについて

電子メールによるWebページの取得と後処理のために次のような2つのツール（mailagent.php, change _path.exe）を作成した。

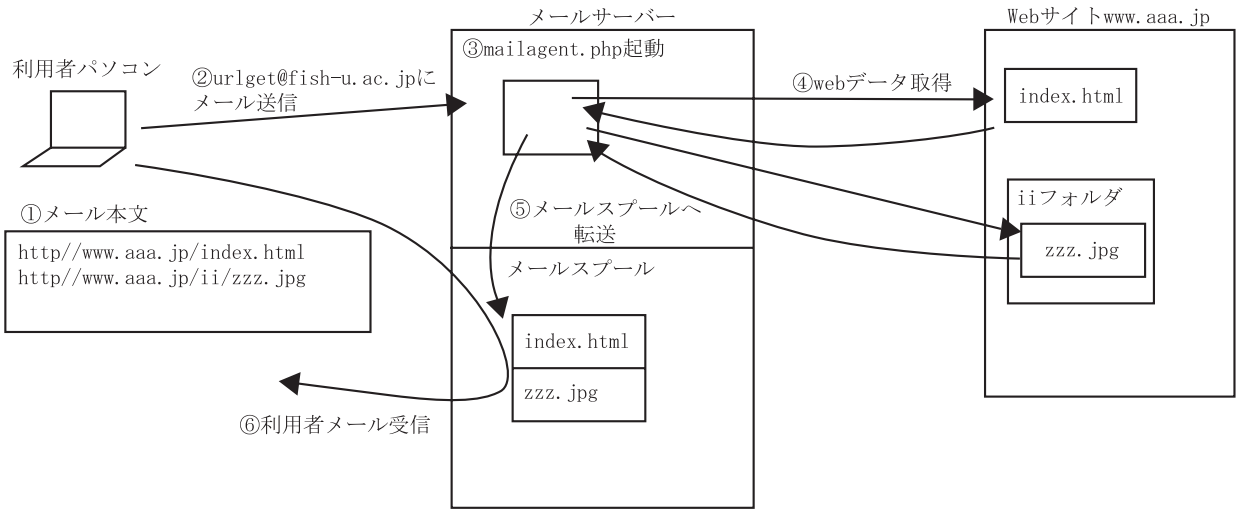
mailagent.php

これは電子メールによってWebページを取得するための主となるツールであり、メールサーバー側で稼動させる。利用者は、メール本文にWebページのURLアドレスを記述し、宛先 urlget@fish-u.ac.jp に発送するとmailagent.phpが実行されるように仕掛けられている。urlget からの返事メールとして、取得したWebデータが戻ってくる。複数のURLを記述でき、個々のURLの内

2006年8月31日受付. Received August 31, 2006.

1 水産大学校水産情報経営学科（Department of Fisheries Information and Manegement）
2 （有）Rapha
†別刷り請求先（Corresponding author）：urikura@fish-u.ac.jp

Fig.1. Role of the tool, mailagent.php



容は個々の添付ファイルとなり戻ってくる。返事メールの本文には、アクセスしたWebページ内にリンクが設定されていれば、そのリンク先URLを取り出し記述するようになっている。

Fig.1.はツールmailagent.phpの働きを示したものである。

①メール本文に取得したいWebページのURLアドレスや画像ファイルのアドレスなどを記述する。
②メールアドレスurlget@fish-u.ac.jpに発送する。
③サーバーのアカウントurlgetにメールが届くと、ツールmailagent.phpが起動する。
④メール本文中のURLアドレスに対してWebデータを取得する。
⑤取得したWebデータは添付ファイルとしてユーザースプールに転送される。
⑥利用者側からのメール受信

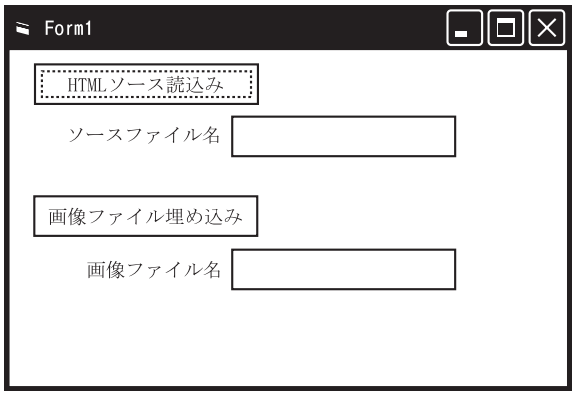
mailagent.phpの実行によって取得できる対象は、Webページの基本的な保存形式であるHTMLファイルをはじめとして、ページに含まれる画像ファイルやPDFファイルなどである。その取得対象のデータのアクセス先であるURLが既知であれば取得できる。

change_path.exe

Windows用に作成されたツールで、その機能は次のようである。閲覧したいWebページの部分が文字のみでよければurlgetによる1回のメール送受信で終了する。しかし、画像などは1回目に取得したHTMLファイルを開いても表示されず、その代り画像の場所に四角の領域のみ表示されるのみである。そこで画像も表示させたいとなれば、再度urlgetで画像データを取得する必要がある。画像データのURLの特定の仕方は後述する。このとき1回目で取得したHTMLファイルと2回目で取得した画像ファイルを同一フォルダに格納したとしても、 HTMLファイルをWebブラウザで開いても画像は表示されない。これはHTMLファイル内の画像指定URLは、画像が用意されているWebサーバー側を指しているためである。change_path.exeは、画像指定URLを単にパソコン側の現在のフォルダだとして書き換える働きをするツールである。

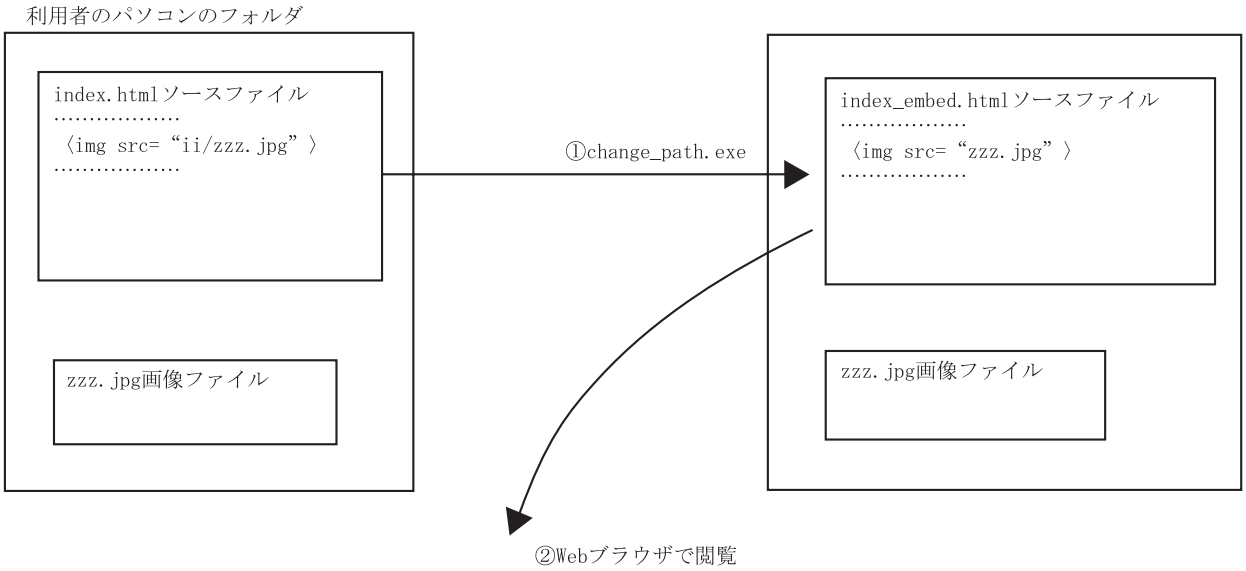
ツールchange_path.exeを起動した画面がFig.2.である。

Fig.2. Dialog box of the tool, change_path.exe



1回目で取得したHTMLファイル名を「ソースファイル名」の入力欄に入力し、「HTMLソース読込み」ボタンを選択する。次いで2回目に取得した画像データのファイル

Fig.3. Role of the tool, change_path.exe



名を「画像ファイル名」の欄に入力し、「画像ファイル埋め込み」ボタンを選択する。これらにより画像ファイルの場所が現在の場所に書き換えられ、HTMLファイル名の後に「_embed」が付加されて保存されることになっている。

ツールchange_path.exeによる作業を図示するとFig. 3.のようである。

状況は、mailagent.phpを使って取得したHTMLソースファイルと画像ファイルが利用者パソコンの同一フォルダに格納されているとしてツールが実行される。

①change_path.exeを起動し、必要なパラメータを入力する。結果は、HTMLソースファイルの画像パス名が書き換わる。

②画像のパス名が書き換えられたファイルをWebブラウザで開く。

3　画像データURLなどの特定

HTMLソース内で、画像ファイルやリンク先のHTMLファイルの場所を指定する方法はいくつかある。これらはHTMLソースを調べれば分かることであるが、ここでは別の方法でこれらのファイルの場所を特定してみる。

1回目で取得したWebページのHTMLファイルをWebブラウザで開き、画像データが必要な場合やリンクを辿る場合など、その該当する箇所のプロパティを表示させる。プロパティ表示の実行は、該当する場所の上でマウス右ボタンクリックで示されるメニューから行える。ここでは画像ファイルのURLを特定することを例示しながら、いくつかのパターンを説明してみる。リンクされたHTMLファイルやPDFファイルであっても同様の方法で場所が特定できる。

HTMLページの場所をソースURLと呼び、仮に

http://www.aaa.jp/bb/cc/dd/eee.html　(1)

とする。サーバーの場所を指すところまでをベースURLと呼ぶと、

http://www.aaa.jp　(2)

である。ソースURLから取得したHTMLソースファイルをパソコンの次の場所（フォルダ）

D:/urlget/html_src　(3)

配下に格納しているとする。画像ファイル名をzzz.jpgとしておく。

例題1　画像プロパティが

file:///D:/xx/yy/zzz.jpg

のような表示の場合、画像URLは(1)のベースURLに画像プロパティで示されたパスを連ねた

http://www.aaa.jp/xx/yy/zzz.jpg

となる。

例題2　画像プロパティが、(3)で示されるHTMLファイルの格納場所から始まる

file:///D:/urlget/html_src/ii/jj/zzz.jpg

のような表示の場合、画像URLは(1)のソースURLに画像プロパティのパスを連ねた

http://www.aaa.jp/bb/cc/dd/ii/jj/zzz.jpg
となる。

例題 3　画像プロパティが、(3)で示されるパスを上位に移動したところから始まる場合で、たとえば 1 つ上位に移動した
file:///D:/urlget/ii/jj/zzz.jpg
のような場合、画像 URL は(1)のソース URL を 1 つ上位に戻して画像プロパティのパスを連ねた
http://www.aaa.jp/bb/cc/ii/jj/zzz.jpg
となる。この例では上位へのパス移動は 1 であったが、移動数がさらに増えた場合は(3)のパス数との関係から画像 URL を決めるためカットアンドトライ的になるかもしれない。

例題 4　画像プロパティに直接画像 URL が表示される場合もある。たとえば、画像プロパティが次のような表示
http://fff.ggg.jp/hh/zzz.jpg
であれば、これが画像 URL そのものである。

4　検索サイトの利用

電子メールによって Web ページを取得する方法の応答例として、検索サイトを利用する方法を紹介する。

検索サイトを使った情報検索は日常よく使われることであるが、これは検索サイトが用意する検索用 Web ページ（フォーム）の文字入力ボックスに検索文字を入力して検索を開始するやり方である。しかし、本稿では Web 通信を使わないで Web データを取得する方法を紹介しているので、このようなフォームは利用できない。

フォームに対応するポイントは、検索プログラムの URL、検索プログラムを利用するためのパラメータ（引数）、日本語検索キーに対する URL エンコード生成と使用文字コードである。

まず、検索プログラムや使用文字コード、URL エンコードなどについての説明は文献[3]などが参考になる。あるいは、前準備として実際に検索サイトを使ってみて、Web ブラウザのアドレス欄に検索プログラムの URL から始まる一群のパラメータや URL エンコードされた文字などが表示されるので、これらから記録しておくことである。

また検索キーに日本語を使う場合、検索プログラムの使用文字コードに合わせた URL エンコード文字列を使う必要がある。この URL エンコード処理を行う Windows 用のフリーソフトとして、たとえば多機能ではあるが WebExe（ウェブエグゼ）[4]が有用である。

次の例題で具体例を示してみる。

例題 5　検索サイト Google について、検索キーとして「漁獲量　比較」で、1 ページ当たりの表示数を 50、PDF 形式のファイルのみ検索したい場合であれば、そのときの検索アドレスのパラメータの決め方は次のようである。

(1) 検索プログラムの URL：
http://www.google.co.jp/search
(2) 検索キーはパラメータ q に代入する。検索プログラムの URL と q の間は記号「?」でつなぐ。
(3) 文字コードは、UTF-8 である。
(4) WebExe で検索キー「漁獲量　比較」を UTF-8 で URL エンコードすれば、
%E6%BC%81%E7%8D%B2%E9%87%8F%E3%80%80%E6%AF%94%E8%BC%83
となる。中ほどにある「%E3%80%80」は全角文字の空白である。他方、半角文字の空白は記号「+」が使われる。
(5) google 検索プログラムのパラメータのいくつかを紹介すると次のようである。
num：検索結果の 1 ページ当たりの表示個数（既定値は 10）
－：キーワードを含まない検索は、そのキーワードの前に「－」記号を付ける
filetype：ファイル形式を指定して検索する場合。ファイル形式としては、
pdf：Abobe Acrobat PDF, ps：Adobe Postscript, doc：Microsoft Word
xls：Microsoft Excel, ppt：Microsoft Powerpoint, rtf：Rich Text Format
などがある。書式は、「filetype：ファイル形式」である。たとえば、PDF 形式を指定したい場合は、「filetype%3Apdf」となる。「%3A」は記号「:」の URL エンコードである。
filetype 以外にもさらにいくつかの検索キーワードが用意されており、検索の効率化を図ることができる[5]。
(6) いくつかのパラメータを使った場合は、パラメータ間を記号「&」でつなぐ。

そこで、冒頭で述べた条件で検索するときのアドレスは次

のようになる。

http://www. google. co. jp/search?q=%E6%BC%81%E7%8D%B2%E9%87%8F%E3%80%80%E6%AF%94%E8%BC%83+filetype%3Apdf &num=50

この検索アドレスを電子メール本文に記述して、宛先 urlget@fish-u.ac.jp に発送すればよい。検索結果は添付ファイルとして戻ってくる。そして添付ファイルを開き、リンクで用意されている Web サイトの URL を特定して、次の Web 情報の取得と続くことになる。

5　手順のまとめ

手順をまとめる前に、まず HTML ファイルの文字コード変換のことに触れておく。

Windows 環境での後処理を想定しているので、ツール change_path.exe を使う必要がある場合、HTML ファイルで使われている文字コード変換が必要な場合がある。HTML ファイルが Windows 標準の文字コード Shift_JIS で記述されているのあればコード変換は必要ないが、UNIX 系の文字コード EUC で記述されている場合、Shift_JIS へのコード変換とが必要となる。このようなコード変換ツールソフトとして、たとえば lfeuc.exe[6]がある。

A.　Web ページの文字情報のみ必要な場合の手順

(1) 電子メールにより HTML ファイルの取得。

B.　Web ページの文字情報に加えてさらに画像が必要な場合の手順

(1) 電子メールにより HTML ファイルの取得。
(2) HTML ファイルを、たとえば Windows のエディタ「メモ帳」で開いたとき、文字化けがあれば、Shift_JIS への文字コード変換が必要。
(3) 第 3 章で述べた方法で、画像データの URL を特定。
(4) 電子メールにより画像データの取得。
(5) 後処理用のツール change_path.exe を使って画像ファイルのパスを変更。

6　あとがき

インターネット接続において、ブロードバンドが利用できないような厳しい利用環境の下、Web ページを電子メールによって取得するためのツールと後処理のツールを作成し、これらの利用法を示した。このような厳しい利用環境の一例として、本校の練習船が航海中、通信手段として人工衛星を経由した電話回線を利用することになる。当然、利用料金は割高となり、通信速度も遅いという環境となる。しかし、Web で提供される情報が必要である状況が生じた場合、ここで述べたツールを使うことが Web データ取得の有効な方法の一つとなる。

ただ、Web ページ取得に当たっての数量的な把握、たとえば衛星回線を使って直接 Web 接続をした料金、あるいは本稿で述べた電子メールを使って Web データを取得する場合の料金など、比較するためのデータは取っていない。これらについては機会があればデータを取得し、別の報告で述べることにする。

最後になりましたが、ツールの検証を天鷹丸一等航海士秦一浩助教授にしていただきました。ここに感謝致します

参考文献

1）総務省編、平成 18 年版情報通信白書、株式会社ぎょうせい、第 1 章 第 2 節（2006）
2）KDDI インマルサットサービス、http://www. kddi. com/business/service/other/inmarsat/index. html
3）URL エンコードと検索エンジン http://www. bousaid. que. jp/software/urlencode/index. php
4）コムシェア http://www.comshare.co.jp/webexe/
5）MB COMPANY、Google グーグルキーワードブック、株式会社宙（おおぞら）出版、135-138（2005）
6）DEKO、http://www.vector. co.jp/vpack/browse/person/an021671.html

付　　録

（1） 後処理を行うパソコンについて

HTML ファイルを開く場合、あるいはツール change_path.exe を使った処理を行う場合は、インターネット接続を行わない状態のパソコンで行うことを薦めます。広告サイトなどが含まれていると、ファイルを開く際インターネットアクセスが実行されてしまうことがあるためです。

（2） URL の記述について

電子メール本文に URL を記述するとき、長い URL は適当な場所で改行を行い複数行にまたがっても１つの URL と判断できるようになっている。また、URL は行頭から書き始めることにすれば、本文中に複数の URL を記述しても対応できるようになっている。