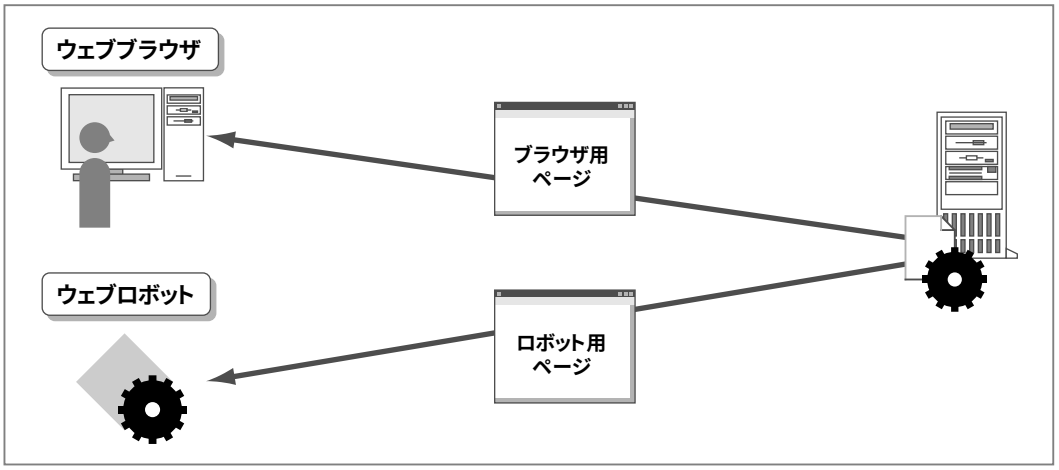
■6-5　動的ページにおけるウェブロボット対策

CGIなどのプログラムを使ってウェブページを動的に生成している際には，これまでみてきたもの以外にも，よりロボットとうまく付き合うためのテクニックがいくつか存在します．続いてはそんなテクニックをみていきます．

ロボットを見分ける

CGIなどのプログラムを使う場合，環境変数を調べることで，アクセスがあったときに，それがブラウザからのものか，ロボットによるものかを調べることができます．既に述べたように，ロボットであるかの判断は，ユーザーエージェント情報やリファラー情報などで行うことができます．したがって，これらの環境変数を調べて，ブラウザからのアクセスとロボットからのアクセスで，異なるデータを送信する，といったことも可能になります(**図6-34**)．

図6-34　ロボットとブラウザで異なるデータを送信する



プログラムによるロボットとブラウザの振り分けを行っている例としては，E-Mail Protectorという Perlのプログラムなどがあります．

- **E-Mail Protector**
  http://www.siteware.ch/webresources/scripts/perl/emp.html

これは非常に単純なPerlのプログラムで，SSIとして利用するものです．ウェブページにメールアド

レスは掲載したいけれども，E-Mailコレクターには回収されたくない，という場合にE-Mailを直接書く代わりに呼び出して使います．

E-Mail Protectorのプログラムは，HTTP_USER_AGENTに入っている名前が，代表的な4つのE-Mailコレクターの名前と同じかどうかをチェックして，該当しない場合は正しいメールアドレスを，該当した場合は嘘のメールアドレスを，なんと10000個も表示します．E-Mailコレクターは不達のメールアドレスを大量に回収してしまうわけで，E-Mailコレクターに対するいやがらせ的な意味ももっているものです(ただ，不達のメールを増やすことは，spam業者がどこかのサーバを第三者中継として利用していた場合，そのサーバに負担をかけてしまうこともありえるので，あまりよいことではありません)．

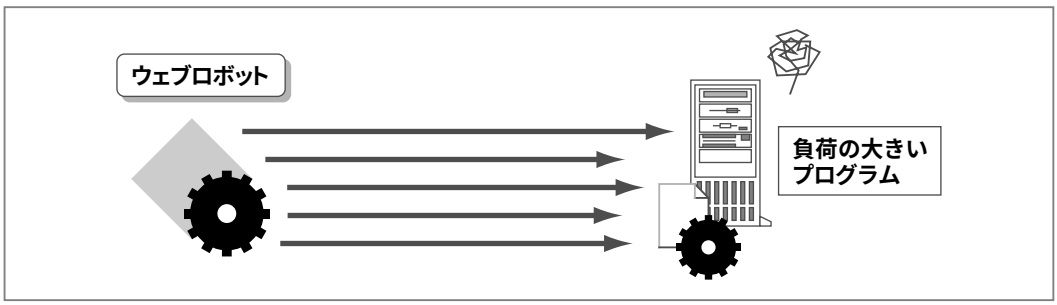
そのうえ，最近のE-Mailコレクターの多くは，Internet Explorerなどと見分けのつかないユーザーエージェント情報を送ってくるものも多く，このスクリプトがそのまま利用できるとは限らないのですが，こういった使いかたもできる，ということはわかるかと思います．

なお，検索エンジンで，自分のページがヒットする確率を上げるために，このような手法をとる場合もありますが，それは，検索エンジンを管理している人たちにとっては迷惑な行為とされています(ロボットが回収したページと，検索結果から利用者が訪れるページが異なってしまっては，検索エンジンの意味がなくなってしまいます)．コラムでも触れていますが，こういったことをしていると，検索エンジン側でページ自体をブロックしてしまって，検索結果に表示されなくなってしまったりして逆効果になることもあるので，お薦めできません．

URLを工夫する

ロボットの中には，動的に生成されたページへのリンクをチェックしないものも存在します．Googleのロボットも以前は動的に生成されたページにはアクセスしないようになっていました．その理由はロボットによってさまざまですが，パラメータがすこし違うだけのURLを大量にインデックスしなければならなくなったり(スパイダートラップと呼ぶ)，ページ生成に利用されるCGIなどはサーバのマシンパワーを多く使用するものも多く，ロボットが大量のアクセスを行うことで，サーバが過負荷の状態になってしまう危険性があるため，というのが主な理由です(**図6-35**)．

図6-35 動的に生成されたページへのアクセスが集中するとサーバがダウンする可能性がある



しかし，ウェブサイト全体を動的なページで構成しているサイトでは，そのせいでまったくロボットからのアクセスが発生しない可能性があります．例えば，大手オンラインブックショップサイトのAmazonでは，トップページ(http://www.amazon.co.jp)にアクセスすると，すぐに動的に生成されたページに転送され，その後もずっと動的なページが続きます．そのおかげで，アクセスした利用者に合わせた情報が表示されるので非常に便利なのですが，動的ページであるため，ロボットに無視されてしまう可能性があります．それが検索エンジンのロボットであれば，そのページは登録されないことになります．Amazonのような商用サイトでは，検索エンジンに自分のサイトが登録されないことは，致命的なことです．そこで，Amazonではちょっとした工夫をしています．URLを以下のような感じにしているのです．

http://www.amazon.co.jp/exec/obidos/tg/browse/-/489986/250-5728042-4552214

つまり，ロボットはURLがCGIなのかどうかを，「.cgi」といった拡張子や，「？」や「&」などの存在でチェックするので，それらをURLから排除しているのです．こうすることで，ロボットはそのページが静的なページであるとみなし，データを収集してくれるようになるわけです．

このようなことを実現するには，Apacheであればロボットの排除のところで紹介したmod_rewriteを使うことができます．簡単なアクセスファイルのサンプルを**図6-36**に示します．

図6-36 mod_rewriteでCGIのURLを静的ファイルのようにするアクセスファイル

```
RewriteEngine on
RewriteBase /
RewriteRule ^view/([0-9]+)/([0-9]+)$ view.cgi?type=$1&id=$2   [T=application/
x-httpd-cgi,L]
```

このサンプルは，以下のように「/」で区切られたURLを分解して，CGIに渡すようになっています．

http://www.example.co.jp/view/12/34567

▼

http://www.example.co.jp/view.cgi?type=12&id=34567

書き換えの際に，MIMEタイプのCGIをあらわす「application/x-httpd-cgi」に変更している点にも注意してください．

■6-6　RSSで情報を提供する

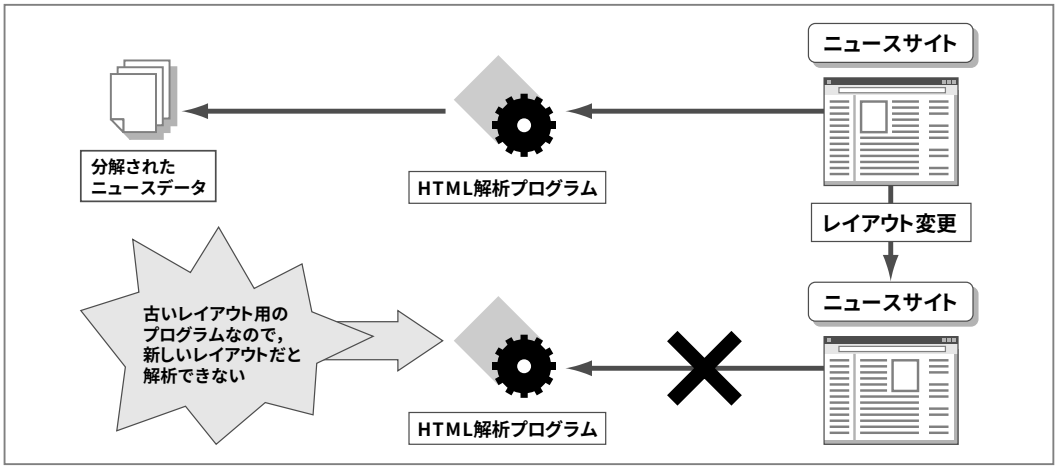
ウェブサイトの情報などをロボットに知らせる手段として，よく利用されるのがXMLです．そこで続いては，RSSというXMLベースのフォーマットを使ってデータの提供を行う方法を紹介します．

サイト情報を提供することの重要性

ウェブ上の株価情報や，ニュース速報，天気予報などのデータを取得して，内容を加工し，表示してくれるソフトウェアがあります．これらは「専用ブラウザ」，「単機能ブラウザ」と呼ばれたりしますが，1つの目的に特化しているため，その情報を必要とする人には非常に便利なツールとなります．

これらのソフトウェアは大抵，ウェブ上で公開されている，特定の情報が掲載されているページのHTMLデータを取得しています．しかし，ロボット用リンク一覧ページのところでも述べたとおり，ウェブページには広告や，他のページへや画像データへのリンクなど，単に情報取得をするためだけであれば不必要なデータがたくさん埋め込まれており，また必要としている情報も，テーブル構造になっていたり，重要な情報が画像になっていたりと，さまざまな表現方法で記述されてしまっています．しかも，HTMLでは，画面上にどうやって表示するか，という情報は埋め込まれていますが，表示されるそれぞれの文字列が，一体どんな意味を持つのか，といった情報はまったくありません．そのため，あらかじめ解析しておいた法則に基づいて，必要な部分だけを切り出すプログラムを作成しておき，そのプログラムを使って，必要な情報を必要な形に変換する，という作業が必要となってしまいます．しかもこの方法は，取得先のウェブサイトが書式を変更すると，解析プログラムもそれにあわせて書き直さなければならないという，大きな問題を抱えてしまっているのです(**図6-37**)．

図6-37　HTML**ファイルの解析プログラムはページレイアウトが変わると作り直しになってしまう**

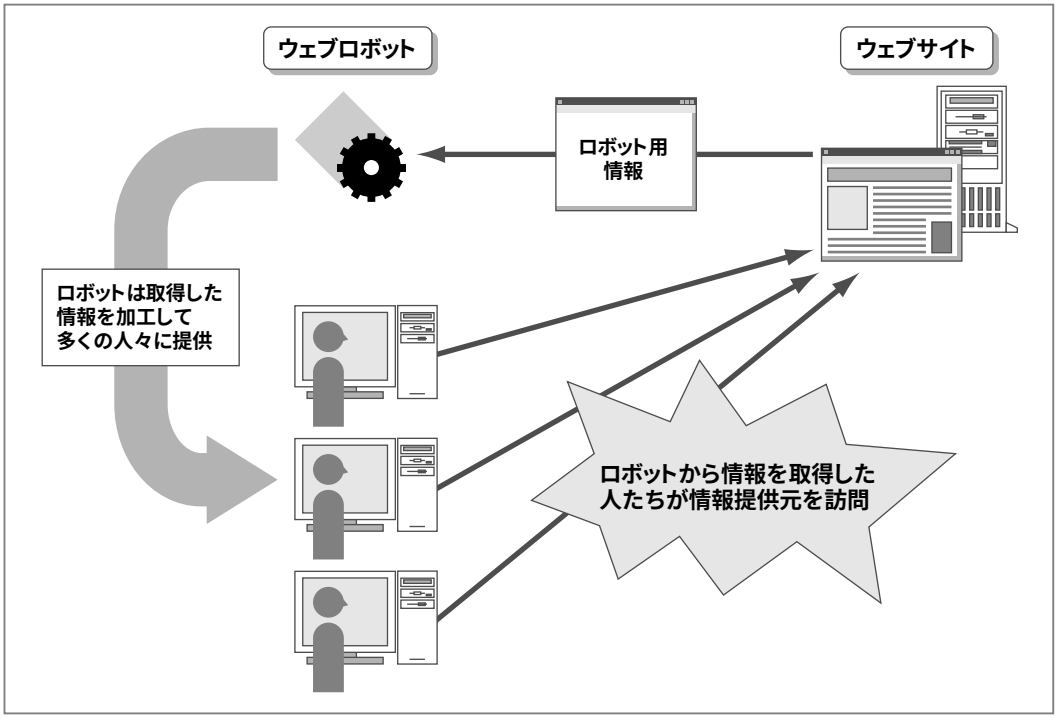


実際，私の友人の田中久太郎氏[*11]は，以前CNNの為替レートのサイトにアクセスして，その時点での為替レートを取得するソフトウェアを公開していましたが，CNNが大規模なサイトデザイン変更を行った際に，それに対応しきれずに，公開を中止してしまいました．

もちろんサイトによっては，ページに貼り付けた広告から収入を得ていたり，著作権の問題などによって，ブラウザを使ってサイトを直接見てもらったりする以外の方法では，情報を提供できない場合もあります．しかし，そういった問題が無く，サイトで公開する情報を，どんどん活用してもらいたいなら，もっとロボットに優しいデータとして，サイトの情報を公開したほうがよいわけです．

ロボットにも使いやすい，ということが有名になれば，よりデータが活用されるようになり，サイトの訪問者も増えるでしょう．例えば，ニュースサイトなどで，見出し情報とURLをセットにしてロボット向けに配信して，その内容を見たいときは，自分のサイトにアクセスしてもらう，という方法を取れば，さまざまなロボットがその情報を取得することで，自分の管理しているサイト以外でも，その情報が公開され，自分のサイトのページへのリンクが貼られるかもしれません．そうすれば，自分のサイトの訪問者も増えてくれるはずです(**図6-38**)．

図6-38 ロボット向けの情報を公開することで，サイトの訪問者も増える



●ロボットが情報を取得しやすいフォーマット

ロボット向けの情報を公開するために，XMLというフォーマットがよく利用されています．この，

XMLとはどんなフォーマットなのでしょうか.

XMLは，HTMLとよく似た，タグを文字列中に埋め込んだ形式になっています．HTMLと同様，タグで囲まれた領域が，エレメント(要素)と呼ばれます．しかしHTMLが，ブラウザで人間が判りやすいよう表示するための情報，例えばページのタイトルや，文字の色や大きさ，表組みなどの情報を埋め込むフォーマットであるのにたいして，XMLは中に書かれているのが，どんなデータであるのか，という情報を埋め込むことができるフォーマットです.

XMLでは，自分が扱いたい情報にあわせて，新しいエレメントとタグを定義できます．例えば，**図6-39**は書籍の情報をXMLで表したものです.

図6-39 書籍情報を表現したXMLのサンプル

```
<書籍情報>
  <ISBN>4-7561-4026-2</ISBN>
  <書名>全速力人生</書名>
  <著者>山田太郎</著者>
  <本体価>2580</本体価>
</書籍情報>
```

書名，作者，ISBNコード，といったエレメントを定義し，タグを埋め込んであるのがわかります．しかもこれらのタグは，HTMLのように画面表示のための情報ではなく，「書名」，「著者」などのデータの意味を表しているので，プログラムで簡単にデータを分解できるわけです.

RSSでサイトの情報を公開する

XMLを使ったデータ配信方法として，最近特に注目されているのがRSSです．これは，ニュースサイトのヘッドラインなどの情報を配布するためのフォーマットです．RSSはRDF Site Summaryの略で，これはつまりRDFというフォーマットを利用して，サイトの要約情報を配信するためのフォーマット，という意味です.

すこしややこしいのですが，まず始めにXMLを利用したRDFというフォーマットがあり，それを利用してサイトの情報を配信しようと作られたのが，RSSなのです．RDFは，W3C[*12]が策定したXMLベースのフォーマットです．まずは**図6-40**に，RDFで書かれたデータのサンプルを示します.

*11：http://www02.so-net.ne.jp/~tanaq/

*12：HTML**のフォーマットをはじめウェブに関するさまざまな規定や指針を策定している団体**

図6-40 RDFのデータサンプル

```
<?xml version=“1.0” encoding=“utf-8”>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
     <RDF:Description href=“http://www.takaaki.info”>
           <dc:title>takaaki.info</dc:title>
           <dc:creator>Mizuno, Takaaki</dc:creator>
     </RDF:Description>
</RDF:RDF>
```

W3Cの文書[*13]には，RDFとはResource Description Frameworkの略で，XMLをベースとした，メタデータをコンピュータにわかりやすい形で記述するためのデータ記述方法，と書かれています．メタデータとは，「データをあらわすデータ」という説明がなされますが，つまりは本の目録のように，何らかのデータに関する種類や内容などの説明データのことです．RDFでは，あらわされるデータ(つまり，先ほどの本の目録の例でいうならば「本」)はURIで示されます．**図6-40**の例は，「http://www.takaaki.info」というサイトに関するメタデータをあらわしており，それを説明する情報とはタイトル(dc:title)と作者(dc:creator)です．

●RSSはどう便利なのか？

RDFを使うことで，さまざまなメタデータをよりロボットが理解しやすい形で提供することができるようになります．そしてそのRDFを使って，ウェブページの更新情報や，現在の内容の要約などを配信できるようにしたのがRSSです．

RSSは現在，多くのニュースサイトや，ウェブログ[*14]などで，公開されている記事のタイトルや要約などを配信することに使われています．配信されたRSSデータは，別のページに組み込

図6-41 RSSリーダーのひとつFeedReader (http://www.feedreader.com/)

んだり，専用のRSSリーダーで読み込んで利用したりすることが可能です(**図6-41**)．

このようなRSSリーダーを利用することで，いちいちブラウザでサイトにアクセスしなくても，新しい記事が公開されたか，そしてその記事は自分に興味のあるものか，といったことを知ることができるようになるのです．あなたがもし，ニュースサイトや，その他定期的に新しい情報を配信するサイトを運営しているのなら，RSSを使ってヘッドライン情報を配信することで，サイト利用者が，適切なタイミングでサイトにやってくることができるよう，手助けすることができるわけです．

さてRSSは，もともとはNetscape社が自社のポータルにおいて，My Netscape[*15]という，さまざまなウェブサイトの情報を集めて，一度に表示してくれるサービスを実現するために策定したものでした(この，Ntescape社が定めたフォーマットはRSS 0.9[*16]と呼ばれています)．RSSはその便利さから，その後Netscape社を離れて進化を続けましたが，その過程で分化が起こってしまい，現在RSSは0.9系，1.0系という2つの系統が利用されるに至っています．

●RSS 0.91

RSS 0.91は，1999年7月に策定された仕様で，Ntescape社のRSS 9.0と，ScriptingNewsフォーマットというUserland社のDave Winer氏によって作られた仕様から作られました．RSS 0.91は今も非常によく利用されています．

- **RSS 0.91 Spec, revision 3**
  http://my.netscape.com/publish/formats/rss-spec-0.91.html

RSS 0.91のサンプルを**図6-42**に示します．

図6-42 RSS 0.91**のサンプル**

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="0.91">
  <channel>
    <title>takaaki.info</title>
    <link>http://www.takaaki.info/</link>
    <description>水野貴明の日々の雑感</description>
    <item>
```

*13：http://www.w3.org/RDF/

*14：blog**とも呼ばれる．読んだ人がコメントを書いたり，自分のウェブログへのリンクを生成させたりできる日記のようなシステム．**

*15：http://my.netscape.com/

*16：http://my.netscape.com/publish/formats/rss-0.9.dtd