i-mode クローラ開発と収集ページ解析

6Y－05

松田 勝志　福島 俊一
NEC インターネットシステム研究所

1 はじめに

携帯電話や PHS などのモバイル端末による WWW(World Wide Web)の利用が盛んである．最も普及しているｉモードでは，2001 年 12 月の時点で契約者数が 2,953 万人以上[1]，ｉモード端末向けの WWW サイトは，2001 年 12 月の時点で 51,000 サイト以上[2]にも上る．そしてこれらのサイトが公開している Web ページはサイト数よりはるかに多いはずである．Web ページ数が増加するに応じて，yahoo! JAPAN[3]のようなディレクトリ型検索エンジンから goo[4]や google[5]のようなロボット型検索エンジンに検索エンジンのパラダイムシフトが起こったように，ｉモード端末向けの検索エンジンも同様なパラダイムシフトが起こることが予想できる．

ｉモード端末向けの検索エンジンの開発には，ｉモード端末向けの Web ページを選択的に収集するクローラの開発が必要である．ｉモード端末向けの google では，ｉモード端末向けではない(PC 向けと呼ぶ)Web ページをｉモード端末で閲覧可能な形式に変換して全文検索システムを提供しているが，ｉモード端末では内容を把握しずらく，必要な情報に辿り着くまでには多くの操作が必要である．また，iYappo[6]では，クローラにより 11 万以上の Web ページを収集している(2000 年 10 月現在)が，OH!NEW?[2]の 51,000 サイトという登録サイト数に比較して規模が小さい．そこで筆者らは，大規模にｉモード端末向け Web ページを収集するクローラを開発した．

本稿では，開発したｉモードクローラの処理方法と，収集したｉモード端末向け Web ページの解析結果を報告する．

2 ｉモードクローラ

ｉモード端末向けの Web ページを選択的に収集するクローラを開発した．基本的にｉモード端末向けであろうと PC 向けであろうと WWW のプロトコル(http)的には差異はない．そのため，従来のクローラでもｉモード端末向けの Web ページを収集することは可能である．

しかし，クローラは単純にリンク先の Web ページを次々と収集するため，クローラの開始 URL がｉモード端末向けの Web ページであってもリンクを辿ると PC 向けの Web ページに到達する可能性がある．例えば，同一 URL でもブラウザの UserAgent に応じてｉモード端末向けと PC 向けの Web ページを切り替えたり，ｉモード端末向けの Web ページ上で PC 向けの Web ページへのリンクを張ったり，等である．そのため，PC 向けの Web ページは収集しないように工夫する必要がある．

本稿で紹介するｉモードクローラは，収集した Web ページを解析し，ｉモード端末向け Web ページか PC 向けの Web ページかを判定し，ｉモード端末向け Web ページである限りリンクを辿って収集するというアルゴリズムを用いた．開発したｉモードクローラはモバイルサーチエンジン WithAir[7]のクローラとして用いている．

2.1 ページタイプ判定

収集した Web ページがｉモード端末向けであるか PC 向けであるかを判定するために，ページタイプ判定[8]を用いている．ページタイプとは，Web ページの構造的特徴で分類した文書の種類であり，例えば「カタログ」や「リンク集」や「求人情報」等がある．ページタイプ判定エンジンは，ある Web ページの構造化文書(HTML)を入力として，その HTML がそれぞれのページタイプにどれぐらい適合しているかを数値として出力する．このページタイプとして「ｉモード」を定義した．図１に「ｉモード」ページタイプの特徴記述の例を示す．

```
加点の条件
    絵文字が使われている
    URL に" /i-mode/" というパスが使われている
    <a href=tel:*>が使われている
減点の条件
    SJIS 以外の漢字コードが使われている
    <frame>が使われている
    ページの総バイト数が 5k バイト以上ある
```

図１「ｉモード」特徴記述

特徴記述は実際にはルールの形で記述されており，図1からも分かるように構造化文書のテキスト本文だけでなく様々な情報を元にある Web ページのあるページタイプらしさ(ここではｉモード端末向け Web ページらしさ)を数値

Development of an i-mode Crawler and Analysis of the collected Web Pages.

Katsushi MATSUDA and Toshikazu FUKUSHIMA
Internet Systems Laboratories, NEC
8916-47, Takayama-cho, Ikoma, Nara, 630-0101 JAPAN

化する.

2.2 精度

ページタイプ判定エンジンを組み込んだ i モードクローラを実装し, 実際に i モード端末向けの Web ページを収集し, その収集精度を測定した. 表 1 に適合率と再現率を示す.

表 1 適合率と再現率

	値	ページ数
適合率	0.996	1246/1251
再現率	0.882	1169/1326

適合率は, i モードクローラが i モード端末向けの Web ページと判定して収集したページ集合を手作業で i モード端末向けかどうかをチェックした. 再現率は, あらかじめ手作業で i モード端末向けの Web ページの集合を定義し, そのページ集合のうち, i モードクローラが i モード端末向けと判定した割合を用いた. 表からも分かるように, 適合率, 再現率ともに非常に高い精度を達成することができた.

3 i モード端末向け Web ページ

PC 向けの Web ページの総数やサイト数の見積もりは様々な文献で公表されている(代表的なものは[9]). しかし, i モード端末向けの Web ページについては OH!NEW?[2]による登録サイト数のみである.

今回開発した i モードクローラは高精度で i モード端末向け Web ページを収集することが可能なため, i モード端末向けの Web ページを広範囲に収集することができた. 今回, 収集したページを元に i モード端末向けの Web ページの総数とサイト数の見積もりを算出した.

3.1 ページ総数

予備収集として, 1,000 ページ程度の典型的な i モード端末向けの Web ページを収集開始ページとして収集を行い, これらのページ集合からハブページ集合を抽出した. ハブページ集合とは, ある Web ページに含まれるリンクのうち, 自身のサイト以外へのリンクの総数が多数あるページの集合であり, ここではリンクの総数が 10 以上とした. これらハブページ集合(18,126 ページ)を種 URL として本収集を行った.

一定期間の収集の結果, ユニークな Web ページとして 4,395,669 ページを収集することができた. 2.2 節で示したように本 i モードクローラの精度は適合率が 0.996, 再現率が 0.882 であるため, 上記の収集ページから想定できるページ総数は, 4,963,816 ページ[1] , すなわち i モード端末向けの Web ページは少なくとも約 5 百万ページはあるということが明らかとなった.

3.2 サイト数

OH!NEW?[2]によると, 2001 年 10 月の段階で i モード端末向けの一般サイトの登録数は 5 万件を超えている. 今回収集したページを元にサイト数を算出した.

URL からサイトを推定することは, 一般的に非常に困難である. ここでは Li らのルールを用いた論理ドメイン抽出[10]に類似した単純なルールによるサイト推定を行なった. ここで用いたルールは, "~"が URL のパスにある場合はその直後までがサイト, "mobile.geocities.co.jp" の場合は 1 つ目のディレクトリまでがサイト, 等の非常に単純なものである. これらのルールを用いて機械的に推定したサイト数は 55,858 サイトであった. サイト推定のもれから, 実際にはこれ以上のサイトがあることが推測できる.

4 おわりに

i モード端末向けの Web ページを選択的に収集する i モードクローラを開発し, その収集精度の評価を行った. また, 収集した Web ページから, i モード端末向けの Web ページは少なくとも 5 百万ページはあることと, 少なくとも 5 万 5 千以上のサイトが現在存在することが判明した.

今後は, 収集した Web ページのリンク構造を調査し, i モード端末向け Web の構造的な解析を行う予定である.

参考文献

[1] DoCoMo Net, http://www.nttdocomo.co.jp/.
[2] OH!NEW?, http://ohnew.co.jp/i/.
[3] yahoo! JAPAN, http://www.yahoo.co.jp/.
[4] goo, http://www.goo.ne.jp/.
[5] google, http://www.google.com/.
[6] iYappo, http://i.yappo.ne.jp/.
[7] 河合ほか, 「モバイルサーチエンジン WithAir の試作と評価」, *情報処理学会研究報告 2001-FI-64*, pp.71-76, 2001.
[8] 松田, 福島, 「文書タイプ分類による問題解決向き WWW 検索システムの開発と評価」, *情報処理学会研究報告 99-FI-53*, pp.9-16, 1999.
[9] Lawrence and Giles, Accessibility of Information on the Web, *Nature*, 400, pp.107-109, 1999.
[10] Li , et al , Constructing Multi-granular and Topic-focused Web Site Maps , *Proceedings of the Tenth International World Wide Web Conference* , pp.343-354, 2001.

1 収集ページ数×適合率÷再現率