決定木による i-mode ページ判定と収集*

6Y－06

釜坂 等, 樋口 毅†
三菱電機株式会社 情報通信システム開発センター§

1. はじめに

WWW の普及に伴い、膨大な情報が発信されており、情報を効果的に抽出し、分類する技術が重要となっている。

本稿では、マイニング手法の一手段である「決定木」を用いてインターネット上の情報を分類する方法の有効性について報告する。

分類の視点として、情報の表示デバイスを取り上げ、i-mode 端末からアクセスすることを意図して作成されたページ（以下 i-mode ページとする）と PC 端末からアクセスされることを意図して作成されたページ（以下 PC ページとする）を意味的解釈なしに分類した。

また、この i-mode ページと PC ページを分類するルール(以下 i-mode ページ判定ルールと呼ぶ）をクローラに組み込み、選択的に i-mode ページを収集する i-mode ページフォーカスドクローラも開発した。

2. i-mode 判定ルールの導出

インターネット上の情報を意味的解釈を行わずに分類するため、説明変数には HTML ファイルのサイズや特定タグの有無等を取り、また目的変数は i-mode ページか否かを取る。説明変数は連続値および離散値であり、目的変数は離散値である事、そして導出したルールが理解しやすい事などから、マイニングツールとして「決定木(C5.0)」を採用した。

i-mode ページを判定する判定ルール導出システムを図 1に示す。

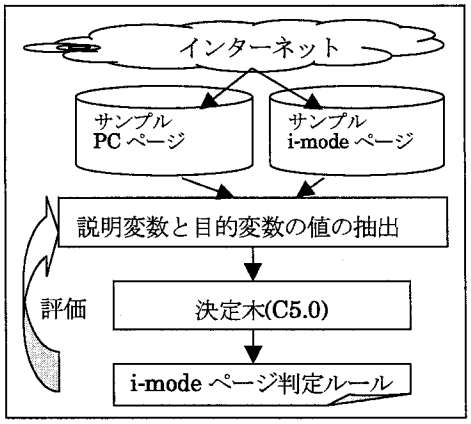

図 1 判定ルール導出システム

まず、インターネットから i-mode ページと PC ページをサンプルとして収集し、目視によって分類を行い、学習用と検証用のデータとした。学習用と検証用のデータとして、それぞれ 2,000 ページ用意した。

説明変数として、PC ページの記述言語である HTML と i-mode ページの記述言語である i モード対応 HTML の仕様にあるタグの種別、タグと属性の組み合わせ、およびタグと属性とその値との組み合わせの値等を用いた。また、i-mode ページの URL の特徴や i-mode ページの特徴も説明変数として加え、最終的に 1580 項目の説明変数を用いた（表 1参照）。

---

* I-mode page detection system by the Decision Tree method.
† Hitoshi Kamasaka, Tsuyoshi Higuchi
§ Mitsubishi Electric Corporation Information & Communication Systems Development Center

i-mode ページ判定ルール導出のパラメータとして、誤判定のペナルティ（例えば、PC ページを i-mode ページと誤判定するよりも i-mode ページを PC ページと誤判定して方がよい等）の設定を行っている。

パラメータの値を変えて複数の i-mode ページ判定ルールを導出し、評価を行った。i-mode ページ判定ルールの中で実際に選択された説明変数は 7 項目から 34 項目であり、<TABLE>、<FRAME> タグ、accesskey 属性の有無やイメージを含むページサイズなどが選択された。

表 1　説明変数の例

説明変数		例
仕様等	タグの有無	<A>,<ABBR>,.. 等
	タグと属性の組合の有無	<A accesskey>,<A href>.... 等
	タグと属性とその値の組合	<A accesskey="0">, <INPUT maxlen="14">,.等
	文字コード	Shift-JIS/JIS/EUC/,半角カナの有無,...等
	絵文字	絵文字の有無,...等
	....	イメージの種別等...
その他	サイズ	テキスト/イメージのサイズ,...等
	URL の keyword	/i/,/i-mode/,...等
	....	....

導出した i-mode ページ判定ルールの適合率(i-mode ページと判定したページの中で実際に i-mode ページであった割合)、再現率(実際の i-mode の中で i-mode ページと判定した割合)の例を以下に示す。

表 2　適合率,再現率

ルール No	説明変数の数	適合率	再現率
ルール 1	23	98%	94%
ルール 2	12	99%	86%
ルール 3	34	96%	96%

3.　判定ルールのクローラへの適用

導出した i-mode ページ判定ルールを Web クローラに適用し i-mode ページを選択的に収集するフォーカスドクローラを作成した。このフォーカスドクローラは、i-mode ページ判定をしながら、i-mode ページのリンクを優先的にクロールする。PC ページへのリンクの場合は優先度を落とし、ある優先度以下になると、それ以降のクロールを行わない。これにより、PC ページの先のリンク先も辿る事により i-mode ページの島（相互にリンクしたページの集合）から PC ページを経由して別の i-mode ページの島へのアクセスおよび収集が可能になる。

図 2に、i-mode ページ収集・判定実験による優先度間の推移割合を示す。例えば、i-mode ページからの 77.9%が i-mode ページへのリンクであり、22.1%が PC ページへのリンクであることを示している。

この結果、PC ページのリンクが 4 段より先のページの収集を止めることにより不要な PC ページの収集をせず、i-mode ページが効率的に収集できることが分かる。

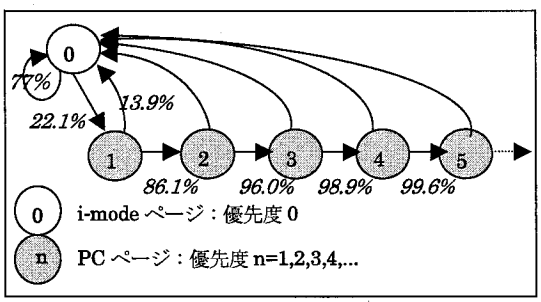

図 2　効率的な i-mode ページ収集

4.　おわりに

本手法を用いる事によって、i-mode ページの意味解釈や深い知見を必要としないで、簡単に i-mode ページ判定ルールが導出できることをを示した。

現在、この i-mode ページフォーカスドクローラにより i-mode ページを約 100 万ページ収集している。実際に収集したページの i-mode ページの適合率と再現率は 99%と 88%で、想定した値とほぼ同じであり、高い i-mode ページ判定精度でかつ効率的に収集することを示した。

また、本手法によれば、定期的にサンプリングとルールの再導出や説明変数の追加を行うだけで、進化しつづける i-mode ページに追従する i-mode ページ判定ルールを導出しかつ収集が可能である。

今後、i-mode ページ以外の Web ページ判定への適用を検討している。