3B1-3

Wrapperを用いたデジタルカタログからの情報抽出

Information Extraction from Digital Catalog Using Wrapper

増山　友美　　松井　藤五郎　　大和田　勇人
Tomomi Masuyama　Tohgoroh Matsui　Hayato Ohwada

*1東京理科大学大学院 理工学研究科 経営工学専攻
Industrial Administration, Faculty of Science and Technology , Tokyo University of Science

*2同 理工学部 経営工学科
Department of Industrial Administration , Faculty of Science and Technology , Tokyo University of Science

We propose a method for extracting information by using Wrapper from the digital catalog distributed with PDF. First, the digital catalog of the PDF is converted into HTML, the HTML is extracted by using wrapper. Second, it paid to attention the problem when PDF convert into HTML and solved by wrapper. Experiments show this method woks well for extraction for PDF.

1. はじめに

近年，Web の急速な発展に伴い，膨大な量の情報が電子化され WWW 上に存在するようになった．個人では，ブログやウェブサイトを立ち上げることにより，自身が持ち合わせている知識・趣味などを公開し他人と共有することができる．また，企業は自らのホームページや Web 広告を用いて商品の宣伝やカタログなどの配布を行い，ショッピングサイトで商品を販売することもできる．こうして我々は，多種多様に存在する情報を様々な形で配信し，パソコンや携帯電話，PDA などの情報端末から場所や時間を選ばず手に入れることができるようになった．

しかしそのような情報があふれる中，逆に必要とする情報だけを得ることは困難となってきている．そのため最近では Web 上から有用な情報を取り出す情報抽出の研究が盛んに行われている．

情報抽出の研究の中でも，Web 上から情報を取り出す研究として Web マイニングが注目を集めている．これは, Web ページで多く用いられてる, HTML や XML の半構造化データから知識を抽出するものである．研究をいくつか挙げると，HTML 形式の類似ページを XML に変換して抽出を行う研究 [2] や半構造化テキストデータからのコンテンツ部分の抽出 [3]，レイアウトパターンによる Web ページ部分情報抽出の研究 [5] などが挙げられる．また Web シラバスからの情報抽出 [6] など特定のデータを対象とした研究も多くされている．これらは定型データを対象とし，文章中で繰り返し出現する文や文の一部のパターンマッチグによりフィールドやレコード単位で情報を抽出する． このパターンマッチングの課題としては，抽出対象ごとにパターンを指定しなければならないことである．例えば，新聞記事ならば，見出し・要約・本文，シラバスなら，科目名，担当教官，概要，単位数などである．これらは，単純な問題を対象とした場合はまだしも，複雑な問題の場合，手作業ですべてパターンを指定することは困難である．本論文では，このようなデータ抽出の中でも，特にデジタルカタログを対象とした抽出方法を提案する．

連絡先: 増山　友美，東京理科大学大学院　理工学研究科　経営工学専攻，千葉県野田市山崎 2641，04-7124-1501，j7408631@ed.noda.tus.ac.jp

2. デジタルカタログ

デジタルカタログとは，インターネットが発達した近年台頭してきたサービスのひとつであり，Web カタログまたはオンラインカタログとも呼ばれる． デジタルカタログは，カタログの中でも特に企業が取り扱う製品一覧カタログを電子化し，Web 上で閲覧またはダウンロードできるようにしたものであり，消費者は，書店でカタログを購入したり，企業にカタログ請求することなくいつでも紙媒体で存在しているカタログと全く同じ情報を手に入れることができる手軽さが普及の一因である．また，企業側としても，印刷データをそのままデジタルカタログへ利用することができるので，カタログの印刷コストや郵送代金の削減ができる．また，ページを拡大できたり，ページをプリントアウトできるなど実際のカタログにはない機能もあり，ネット上で多くの人に見てもらえることから注目を集めている．多くのデジタルカタログは PDF 形式で配布されていることが多い．

```
<HTML><TITLE>Some Country Codes</TITLE><BODY>
<B>Congo</B> <I>242</I><BR>
<B>Egypt</B> <I>20</I><BR>
<B>Belize</B> <I>501</I><BR>
<B>Spain</B> <I>34</I><BR>
</BODY></HTML>
```

(a)

```
procedure ccwrap_LR(page P)
  while there are more occurrences in P of '<B>'
    for each ⟨ℓ_k, r_k⟩ ∈ {('<B>', '</B>'), ('<I>', '</I>')}
      scan in P to next occurrence of ℓ_k; save position as start of kth attribute
      scan in P to next occurrence of r_k; save position as end of kth attribute
  return extracted {..., ⟨country, code⟩, ...} pairs
```

(b)

図 1: LR-Wrapper

しかし，デジタルカタログを PDF 形式で配布する時の欠点として，現状多くのデジタルカタログはすでに発行されているカタログをそのままイメージスキャナで取り込んでいるため様々な属性の情報が掲載されているにもかかわらず．テキスト

検索しかできないことである．そのため，閲覧者が 10,000 円以下のソファを探すまたは，テーブルのみを閲覧したいといったような要望があっても属性をともなう検索が出来ないため，検索することが出来なかった．
そこで本論文では，このデジタルカタログがテキスト検索しか行えないという欠点に着目し，属性検索できるように PDF 形式のデジタルカタログを HTML に変換し Wrapper をもちいて情報抽出を行う．

3. 関連研究

本研究で用いる技術に Wrappper がある．Web Wrapper[8] とは Web ページからある特定の部分を自動的に抽出するためのプログラムのことである. Web Wrapper のプロセスとして

1. Web ページから必要な情報が含まれている情報を特定
2. その特定部分を抽出するための WebWrapper を構築
3. 一度 Wrapper を構築すると次回からは自動で必要な情報を抽出する

が挙げられる．Wrapper で抽出した情報は, 統合・再利用することによって新たな利用価値が生まれる．

Wrapper の研究として, Kushumerick の Wrapper Induction[1] が挙げられる．これは, 機械学習を用い, 訓練例を入力として与えることにより, Wrapper 生成を行うものである．例として Kushumerick が提案した LRwrapper を挙げる．

図 1(a) は, 国名と国コードを表示する HTML である．この HTML 形式から国名と国コードだけを取り出したい場合, 国名なら $<B>$ と $</B>$ の間, 国コードなら $<I>$ と $</I>$ の間に挟まれている単語を抽出すればよい. 以上をアルゴリズムであらわしたものを, 図 1(b) に示す.

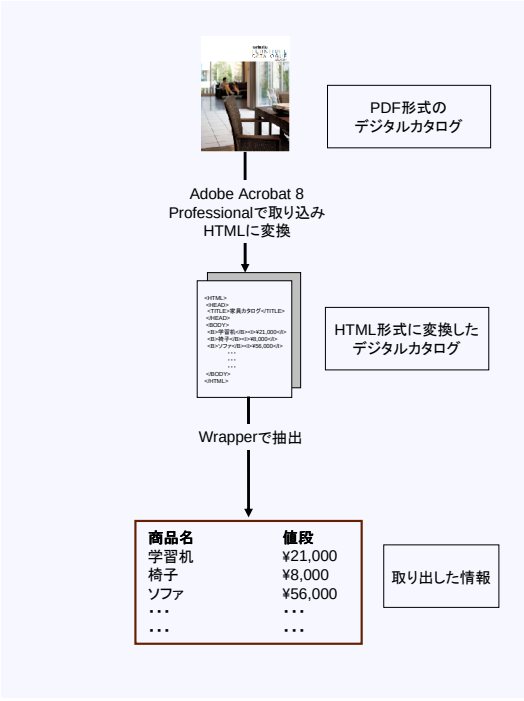

図 2: 提案手法の流れ

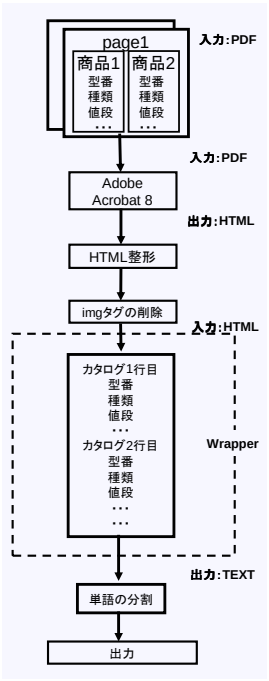

図 3: システム構成

Kushumerick はこのほかに HLRT, OCLR, HOCLRTwrapper やネスト構造でも用いることのできる N-LR や HLRTwrapper を提案している. また, この Kushumerick の研究発展させたものとして, 植松 [4] らの特殊 wrapper や山田 [7] らの PLRwrapper が挙げられる．本論文では，この Wrapper を用いて HTML から情報抽出を行う．

4. 提案手法

提案手法の流れを図 2 に載せる．提案手法では，PDF 形式のデジタルカタログを HTML 形式に変換することで有用な情報を抽出する手法を提案する．この手法では，もともと PDF で存在するデジタルカタログを Adobe System 社が提供する PDF 編集ソフト，Adobe AcrobatR 8 Professional に取り込み，Acrobat の機能の一つである，PDF 文書の HTML 変換機能を用いて PDF 文書を HTML 形式に変換する．そして変換 HTML から必要な情報を Wrapper を用いて抽出することを行う．

しかし，この提案手法で問題となるのが PDF を HTML に変換するときに起こる HTML の変換ミスである．具体的には，

1. 1 つの単語が分割されてタグに囲まれることがある
2. 複数の単語が連結して 1 つのタグで囲まれていることがある

のふたつが起こる場合がある．この問題が起こるとうまく抽出できず，データベースに格納し，情報を再利用するといったことができなくなる．そのため，このふたつの問題を解決するような Wrapper を構築し，抽出を行う．

4.1 抽出対象

本研究におけるデジタルカタログとは，Web 上に PDF として公開している商品カタログで，同様の紙媒体のカタログが存在するカタログを対象とする．また，1 ページに複数の商品情報が掲載されており，抽出する商品は，型番，種類，値段のど

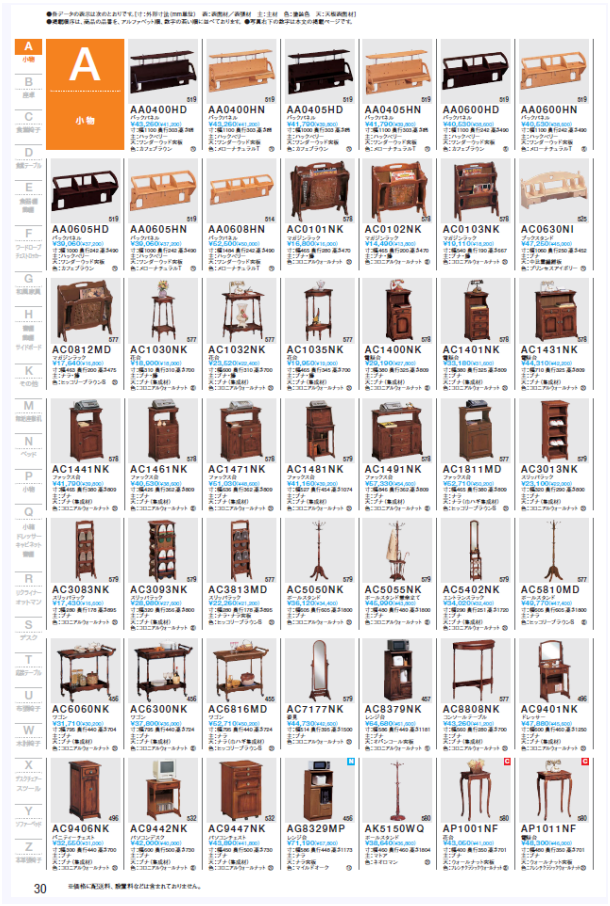

図 4: カタログ例

AA0400HD	AA0400HN	AA0405HD	AA0405HN	AA0600HD	AA0600HN	
バックパネル	バックパネル	バックパネル	バックパネル	バックパネル	バックパネル	
¥43,260 (¥41,200)	¥43,260 (¥41,200)	¥41,790 (¥39,800)	¥41,790 (¥39,800)	¥40,530 (¥38,600)	¥40,530 (¥38,600)	
寸:幅1100 奥行303 高さ608	寸:幅1100 奥行303 高さ608	寸:幅1000 奥行303 高さ608	寸:幅1100 奥行303 高さ608	寸:幅1100 奥行242 高さ490	寸:幅1100 奥行242 高さ490 7672672672672	
主:ハックベリー	主:ハックベリー	主:ハックベリー	主:ハックベリー	主:ハックベリー	主:ハックベリー	
天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	
色:カフェブラウン	色:メローナチュラルT	色:カフェブラウン	色:メローナチュラルT	色:カフェブラウン	色:メローナチュラルT	
AA0605HD	**AA0605HN**	**AA0608HN**	**AC0101NK**	**AC0102NK**	**AC0103NK**	**AC0630NI**
バックパネル	バックパネル	バックパネル	マガジンラック	マガジンラック	マガジンラック	ブックスタンド
¥39,060 (¥37,200)	¥39,060 (¥37,200)	¥52,500 (¥50,000)	¥16,800 (¥16,000)	¥14,490 (¥13,800)	¥19,110 (¥18,200)	¥47,250 (¥45,000)
寸:幅1000 奥行242 高さ490	寸:幅1000 奥行242 高さ490	寸:幅1484 奥行242 高さ490	寸:幅465 奥行280 高さ470	寸:幅465 奥行200 高さ470	寸:幅540 奥行190 高さ667	寸:幅1060 奥行250 高さ452
主:ハックベリー	主:ハックベリー	主:ハックベリー	主:ブナ・籐	主:ブナ・籐	主:ブナ・籐	主:ブナ
天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	色:コロニアルウォールナット	色:コロニアルウォールナット	色:コロニアルウォールナット	天:中比重繊維板
色:カフェブラウン	色:メローナチュラルT	色:メローナチュラルT	色:プリンセスアイボリー			

図 5: 結果出力

AA0605HD	AA0605HN	AA0608HN	AC0101NK	AC0102NK	AC0103NK	AC0630NI
バックパネル	バックパネル	バックパネル	マガジンラック	マガジンラック	マガジンラック	ブックスタンド
¥39,060 (¥37,200)	¥39,060 (¥37,200)	¥52,500 (¥50,000)	¥16,800 (¥16,000)	¥14,490 (¥13,800)	¥19,110 (¥18,200)	¥47,250 (¥45,000)
寸:幅1000 奥行242 高さ490	寸:幅1000 奥行242 高さ490	寸:幅1484 奥行242 高さ490	寸:幅465 奥行280 高さ470	寸:幅465 奥行200 高さ470	寸:幅540 奥行190 高さ667	寸:幅1060 奥行250 高さ452
主:ハックベリー	主:ハックベリー	主:ハックベリー	主:ブナ・籐	主:ブナ・籐	主:ブナ・籐	主:ブナ
天:ワンダーウッド突板	天:ワンダーウッド突板	天:ワンダーウッド突板	色:コロニアルウォールナット	色:コロニアルウォールナット	色:コロニアルウォールナット	天:中比重繊維板
色:カフェブラウン	色:メローナチュラルT	色:メローナチュラルT	色:プリンセスアイボリー			

図 6: 付加情報量の違いによるずれ

の商品にも必ずある必要情報と，寸法，主材，表面材，塗装色などの商品によって情報が異なる付加情報である．今回，商品画像は抽出対象外とし，テキストのみを抽出する．

4.2 システム構成

システム構成図を図 3 に載せる．まず，PDF 形式の商品掲載ページを Acrobat に取り込み HTML で出力し，HTML 整形を行う．

そして不要な img タグを削除し Wrapper にかける．Wrapper では，商品の情報を 1 行ずつ読み込み，1 行分を連結して取り出す．連結することによって，分割された単語も連結して変換された単語も同じ連結された状態になる．結合した単語は，型番はカタログ列数で，製品の種類は辞書一致，値段は正規表現，付加情報は正規表現または辞書一致による分割を行う．

5. 出力結果

本手法に則ってシステムを実装した．データは，カリモク家具カタログ 2006-2007 版の商品一覧 (写真索引) を使用した．商品のカタログ例を図 4 に載せる．

出力例として，30 ページを図 5 に載せる．出力は，上から商品型番，種類，税込価格 (本体価格)，と大きさ，主材などの付加情報で表示される．

抽出では，商品数 48 のうち 28 を情報の欠損なく抽出できた．単語でみると，単語数 325 のうち 304 の単語を抽出できた．抽出の精度は，商品数では 58.3%，単語数の精度は 93.5%であった．

6. 考察

本手法を実装した結果の考察を行う．

6.1 抽出精度

単語における抽出では，今回提案した手法により，高い精度を得ることができた．単語を正しく抽出できなかった理由としては，同じ単語が異なるタグで囲まれている場合があげられる．さらに今回提案した手法でカバーできなかった商品数の精度に対しての考察を行う．

6.2 1 行に 2 段にわたり情報が表示されている場合

商品情報の抽出が正しくできなかった原因のひとつとして，1 行に 2 段にわたり情報が表示されている場合が挙げられる．これは，例えば高さを変えられる商品で 1 行に最大の高さと最小の高さが掲載されていることがあった．この場合，提案手法では 1 行分しか抽出できなかったため，最小の高さのみ抽出された．最大の高さは数字のみ抽出されたが，別の商品の情報に付随して抽出されていたため，どの商品の情報なのかがこの手法では，特定することが困難であった．

6.3 付加情報量が同じ行で異なる場合

商品情報が正しく抽出できなかった原因のふたつ目が付加情報量が行内で異なるときである．この場合，1 行ずつ横に読み込んで HTML 化する本手法では，列ごとに情報量が異なると，空白を読み込めず，取り出した情報がつめられて表示され

てしまう (図 6)．その為, 1 つの商品に，別の商品の属性が現れてしまうことがある．この場合，空白を認識することで属性のずれを解決することが可能である．

7. まとめ

本研究では，PDF 形式のデジタルカタログ HTML に変換し Wrapper を用いることによって有用な情報を抽出する手法を提案した．本提案により，PDF から情報を抽出することができ，PDF から情報の再利用を行うことが可能であることがいえた．しかし，この提案手法は使用するソフトウェアの性能に依存するところもあり，考察で述べたようにまだまだ多くの問題を抱えている．これらの問題解決に関してはまだまだ議論の余地があるといえる．今後の展望としては，今回浮上した問題に対する解決方法の模索と，HTML の整形など手動で行っていた部分を自動化し，より精度の高い抽出を行いたい．また抽出対象をデジタルカタログだけではなく，他の抽出対象でも応用できる Wrapper の構築を行っていきたいと考える．

参考文献

[1] N. Kushmerick Wrapper induction : efficiency and expressiveness, Artificial Intelligence,Vol. 118, No. 1-2, pp. 15–68, (2000).

[2] 板井久美, 高須淳宏, 安達淳, HTML からの情報抽出と統合, NII Journal, No. 6,pp.9–19, (2003).

[3] 池田大輔, 山田泰寛, 廣川佐千男, Web 上の多言語テキストデータからのラッパー自動生成, 情報基盤センター年報, (2003).

[4] 植松幸生, 内山俊郎, 片岡良治, 松井藤五郎, 大和田勇人, 複数の Web Wrapper による高精度な情報抽出, 2007 年度人工知能学会（第 21 回）全国大会講演論文集, 2G4-2 (2007).

[5] 韓浩, 徳田雄洋, Web 部分情報抽出システムとその応用, 日本ソフトウェア科学会第 23 回大会論文集, 4B-1, (2006).

[6] 山田信太郎, 松永吉広, 伊東栄典, 廣川佐千男, Web シラバス情報収集エージェントの試作 (!特集?ソフトウェアエージェントとその応用論文), 電子情報通信学会論文誌.D-I, 情報・システム, I-情報処理, Vol. J86-D-I, No. 8,pp. 566–574, (2003).

[7] 山田泰寛, 池田大輔, 廣川佐千男, 半構造化文書に対する木構造と文字列を組合せたラッパーの自動生成法, 第 72 回情報学基礎研究会, 第 157 回自然言語処理研究会,(2003).

[8] 山田泰寛, 池田大輔, 坂本比呂志, 有村博紀, WWW からの情報抽出: Web ラッパーの自動構築 (特集 WWW 上の情報の知的アクセスのためのテキスト処理), 人工知能学会誌,Vol. 19, No. 3, pp. 302–310, (2004).