Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

EC サイトからの商品情報抽出ルールの自動生成

飯村結香子[†] 真鍋知博[††] 塩原寿子[†] 内山匡[†]

EC サイトの商品説明ページから，商品名や価格情報，商品説明文等の商品情報を自動で抽出する商品情報抽出ルールの自動生成ついて提案する．EC サイトで商品説明ページの定型性を利用し．複数のページ間を比較したときに，共通な部分と変化する部分を分離する．商品属性ごとに変化する部分を抽出位置の候補として，その部分から抽出される値の特徴およびその部分の周辺の固定の特徴からその出現位置候補が商品情報の抽出位置であるか否かを判定する．商品属性ごとに定義された属性値抽出箇所を，その商品説明ページの商品情報抽出ルールとする．提案手法を実装し抽出実験と評価を行い一定の有効性を確認した．

Automatic rule generation for extract information from E-commerce Web Page

Yukako IIMURA[†] Tomohiro MANABE[††] Hisako SHIOHARA[†] and Tadasu UCHIYAMA[†]

We propose an automatic rule generation methods in order to extract commodity information from the E-commerce sites. The commodity explanation pages are composed by both the changing parts and the fixed parts, and the commodity information is expected in the changing parts. Besides, the types of the information such that item name, price, etc. are to be estimated by using the peripheral features. We made a prototype of our method and applied to real E-commerce sites pages, and the experimental results show definite effectiveness of our method.

1. はじめに

E-Commerce サイト(以下，EC サイト)では多くの商品が掲載され販売されている． EC サイトの Web ページから商品情報を抽出し，どのような商品が販売されているか，あるいはユーザの閲覧履歴を利用して閲覧した Web ページから商品情報を抽出し，そのユーザがどのような商品を閲覧したか等をデータ化し，このデータを用いることで種々のマーケティング分析が可能であると考えられる．そこで我々は EC サイトの Web ページからの商品情報の抽出に取り組んでいる．

ここでいう商品情報とは，ある商品に関する商品名や商品価格といった商品属性の集合である．たとえば，ある Web ページに掲載された商品について，商品名「花柄シフォンスカート」，価格「9800 円」，商品説明文「適度な透け感のある素材で，シルエットがきれいです...」等の商品属性の集合が商品情報として抽出される．

各商品が持つ値である「花柄シフォンスカート」や「9,800 円」といった値を商品属性値と呼び，その商品属性値がどのような種別のものであるかを示す「商品名」や「価格」を商品属性名と呼ぶ．

しかし，Web ページは人がブラウザを利用して閲覧，データを認識，理解するという利用方法が前提になっているため，機械的に Web ページから商品情報を抽出することは容易ではない．

Web ページを半構造化文書として扱い，同種の構造を持つ Web ページから特定の情報の抽出するプログラムや，抽出箇所を指定するルールの作成についての種々の研究が行われている．このような抽出プログラムや抽出ルールを Web ラッパーとよぶ．一度 Web ラッパーを生成すると次回からは Web ラッパーを生成するプロセスなしに同種構造を持つページから自動的に情報を抽出することができる．ただし，EC サイトが異なると Web ページの構造が異なるため EC サイトごとに Web ラッパーの生成が必要となる．また，Web ラッパーを生成した後にサイトのデザイン変更が起きると Web ページの構造が変わるため再度 Web ラッパーの生成する必要がある．

Web ラッパーの生成方法には，人手による方法が考えられる．しかし，多くの EC サイトの Web ページから情報の抽出を行いたいことや，EC サイトでデザインの変更が起きるたびに Web ラッパーを生成し直す必要があることからメンテナンスコストが大きくなる．また Web ラッパーの生成者は Web ページの記述に用いられる HTML 言語などを熟知していることが求められる．

このため，機械学習を用いて Web ラッパーを自動で生成する方法も提案されている[2].

[†] 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
[††]京都大学大学院大学院情報学研究科
Graduate School of Informatics Kyoto University

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

しかし，この場合においても学習データとして，抽出したい情報とその抽出箇所のペアからなるデータを Web ページごとに人手で生成する必要がある．GUI を実装することでラッパーの生成や学習データの生成を支援する研究もおこなわれている[3][4]．しかし，情報を抽出したいサイトが増えるたびに，あるいはサイトのデザインが変更されるたびに人手で学習データを生成する必要があるなどコストが大きい．
そこで，商品説明ページからの商品情報抽出ルールを自動で生成することを検討した．商品説明ページを観察すると，以下のような性質を持つことが分かった．1) 同じサイトの商品説明ページであればほぼ同一の構造を持っていること．2)複数の商品説明ページを比較することで，ページ間で共通な部分と，各ページで変化する部分に分けることができ，商品情報は変化する部分に存在すること．
また，変化する部分が商品属性値であるか，どの商品属性に対応するものかについて，変化する部分から取得される値，および変化する部分の周辺の共通部分からある程度推測可能であると考えた．そこで本稿では，商品情報を抽出したい商品説明ページの集合を入力として与えることで，Web ラッパーつまり商品情報抽出ルールを自動生成する方法について提案する．また，提案手法を実装し，商品情報抽出ルールを自動生成し，この商品情報抽出ルールを商品説明ページに適応して商品情報を抽出した結果を評価することで，提案手法の有効性を示した．
本稿の構成は以下のとおりである．まず 2 章で本稿が抽出対象とする商品情報および商品説明ページについて概説し，3 章で商品情報抽出ルールの自動生成法について説明する．4 章で実験と評価について報告する．5 章はまとめと今後の課題である．

2. 商品情報と商品説明ページの特徴について

本章では本稿で生成する商品情報抽出ルールが抽出する商品情報について，また抽出元となる商品説明ページの特徴について述べる．

2.1 商品情報

商品情報とは，商品に関する商品名や商品価格といった商品属性の集合であり，商品属性は商品に関する値である商品属性値とその種別を表す商品属性名からなる．
EC サイトにより記載される商品属性は異なるが，多くの EC サイトで記載されている商品属性には以下のような項目がある．

- 商品名：掲載商品の名称．
- 商品コード：サイト内で商品を識別するためのコード．JAN コードや ISBN など共通コードと EC サイト独自に付与されるコードがある．
- 価格：商品の販売価格，通貨，提示されている価格が税込・税抜どちらであるかなどを含む．
- 商品説明文：商品についての説明文章．
- 商品画像：商品の写真．閲覧するユーザにブラウザ上で表示されるのは，Web ページ内で指定された URL から取得された画像である．
- パンくずリスト：そのサイトが定義する構造において現在表示しているページへ至るパス．その EC サイトの商品カテゴリ階層において，表示している商品が所属する位置が示されることが多い．例) レディスファッション＞スカート

2.2 商品説明ページの特性

多くの EC サイトでは一つの商品説明ページに一つの商品情報を掲載している．一部の商品説明ページでは一つの商品説明ページに複数の商品について掲載しているがこれは全体のごくわずかにすぎない．このため，本稿では一つの商品情報を掲載している商品説明ページのみからの商品情報抽出ルールの自動生成を検討する．
商品説明ページは HTML で記述されるが，商品情報はブラウザで表示されるテキストノードと，ブラウザ上には表示されない hidden 要素などの属性値との双方に記述されている．表示される部分はユーザへの商品情報の提示を目的として記述されており，非表示部分は EC サイト側が利用することを目的として記述されている．
EC サイトでは，同サイトのロゴやサイト名など常に表示したい情報，よく使うメニューなどはヘッダやサイドメニューに，そのページで掲載する情報はメインパートになどレイアウトが統一されている．商品説明ページなど，同種の情報を掲載するページが大量にある場合には，メインパートの中でもどこにどの情報を提示するかのレイアウトもほぼ統一されている．
実際の EC サイトの商品説明ページを観察した結果，次のような特性がみられた．

- 同じ EC サイトの同じ種類のページはほぼ同じ構造を持つ．ページ間で共通な部分はヘッダやフッタおよび，メインパートの定型の文言であり，変化する部分には個別の商品情報や，そのページを閲覧しているユーザに向けた情報が含まれる．
- 商品情報の各商品属性は一定のパターンで記述されており，出現箇所や商品情報前後に現れる文字列が一定である．

上記を「同一種類の商品説明ページにおける定型性」と呼ぶことにする．「同一種類の商品説明ページにおける定型性」をもつ EC サイトの商品説明ページのレイアウトを簡略に示した一例が図 2-1 である．灰色の箇所がページごとに変化する部分であり，それ以外の部分が複数のページ間に共通で現れる部分である．
商品説明ページは，あらかじめページ種類ごとに準備したレイアウト（以下，テンプレートと呼ぶ．）に，個別の商品情報や，その Web ページを閲覧しているユーザに合わせた情報を挿入することで生成していると考えられ，これにより「同一種類の商品説明ページにおける定型性」が生じていると考えられる．ここで，商品情報やユーザの情報が挿入されるページごとに値が変化する部分をスロットと呼ぶことにする．た

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

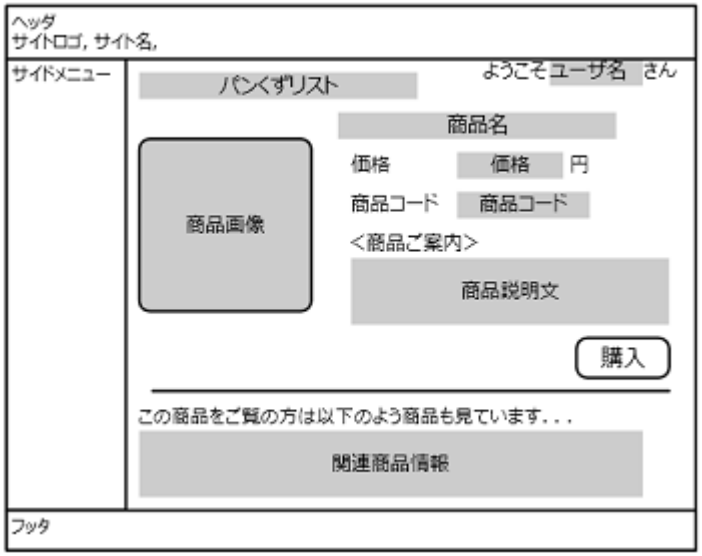

図 2-1. EC サイトの商品説明ページのレイアウトを簡略に示した一例

だし，複数の EC サイト間で成り立つとは限らない．
スロットにどのような種類の情報が入るかは，その周辺にどのような種類の情報が存在し，推測可能な場合とそうでない場合がある．例えば，図の例では，商品コードの属性値が入るスロットの前方に商品コードという文字列が存在し，価格の属性値が入るスロットの後方に通貨記号が存在している．一方で人が見た場合にレイアウトやその属性値から商品属性名が判断できる場合，例えば，商品名や商品説明文などには商品属性名の記載は省略されることも多い．

3. 商品情報抽出ルールの生成

２章で述べた商品情報や EC サイトの商品説明ページで特性から，商品説明ページの集合を観察することにより，テンプレートやスロットを推測することができ，あるスロットから抽出される情報がどの商品属性に対応するかがある程度推測できると考えられる．
このことから以下の手順による商品情報抽出ルールの自動生成を検討する．1) 商品説明ページ群を収集する．2) 1)で集めた商品説明ページを相互比較することにより，テンプレートを解析し，これを利用して商品情報抽出ルールを自動生成する．本稿では，2)の商品情報抽出ルールを自動生成する方法について述べる．

3.1 商品情報抽出ルールの自動生成

商品情報抽出ルールの自動生成は以下のようなプロセスで行う．
まず，複数の商品説明ページを比較して，テンプレート内の固定的に現れる部分を推測する．次に，変化する部分にスロットが存在するとし，変化する部分から取得される値の特徴およびその周辺固定部分の特徴をもとに，変化する部分から商品属性値が抽出されるか否かを判定する．最後に商品属性ごとに定義された商品属性値の抽出位置のペアの集合を商品情報抽出ルールとする．本稿では商品属性値の抽出位置を XML 文書の特定の部分を指し示す構文として規定された XPath[5]を用いて表現することとした．この XPath の記述についての検討についても述べる．
商品説明ページは閲覧者に対して商品の情報を提示することを目的としていることから，抽出すべき商品属性値はブラウザに表示される Web ページのテキストノードに存在すると考えられる．ただし，商品画像については，IMG 要素に記載されている src 属性がブラウザによって解釈され表示される．そこで，本稿では商品属性値の抽出およびを商品属性名の推定はテキストノードと IMG 要素から行うこととする．

3.1.1 テンプレート固定部分の推測

テンプレートで固定的に表れる部分を推測するには，複数の商品説明ページで同一の位置に出現する文字列または IMG 要素の src 属性が共通であるかを調べればよい．ただし，入力として与えられる Web ページの集合に，商品説明ページ以外が混入することなども考えらえるため，全入力 Web ページのうち，ある閾値を超えるページで共通であれば固定部分とする．
ここで出現位置の表現には様々な方式が考えられるが，本稿では出現位置を示すのにルートノードから注目するノードまでのパスの，パス上の要素名を書き並べることによる表現する．これをタグパスと呼ぶことにする．
タグパスとテキストノードから取得される文字列をペアにしたものを用いて処理を行う．ただし，IMG 要素については，タグパスと IMG 要素の src 属性の値から取得される値をペアにしたものを利用する．
例えば図 2. 商品説明ページ HTML の一例での「花柄シフォンワンピース」という文字列は，以下の表現となる．
/html/body/div/h1/text()：花柄シフォンワンピース
テンプレートの固定部分について，あるタグパスで文字列の全体が固定である場合と，あるタグパスで一部の文字列が固定である場合がある．例えば，図 3-1. の「価格」という文字列が複数の文書で共通に出現する場合は，そのタグパスの文字列全体が固定であるが，「8,800 円」の一部「円」だけが共通に出現する場合は，タグパスの文字

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

```
<HTML>
<HEAD>
<TITLE>【Aショップ】花柄シフォンワンピース ｜ファッション</TITLE>
</HEAD>
<BODY>
<DIV>
<H1>花柄シフォンワンピース</H1>
<IMG SRC="/img/item/abd1235.img">
<TABLE>
<TR><TD>価格</TD><TD> 8,800円</TD></TR>
<TR><TD>商品コード</TD><TD> ABD1235</TD></TR>
</TABLE>
<DIV>
<B>＜商品ご案内＞ </B>
これからの季節にピッタリ.
大きな花柄が印象的でシフォンの透け感と色合いが美しいワンピースです.
</DIV>
</DIV>
<DIV>

</DIV>
</BODY>
</HTML>
```

図 3-1. 商品説明ページ HTML の一例

列の一部が固定である．それぞれの抽出方法について述べる．

全体固定部分の抽出

全体固定部分であるか否かは，商品説明ページの集合でのタグパスと文字列のペアの出現率を見る．出現率が閾値を超えればそのペアを全体固定部分であるとして抽出する．

一部固定部分の抽出

一部固定部分については，商品説明ページ集合のあるタグパスのペアとなっている文字列の共通部分文字列を求め，その出現確率を見る．出現率が閾値を超えれば，そのタグパスと共通部分文字列を一部固定部分であるとして抽出する．共通部分文字列には, 文字列の前方が一致する (プレフィックス), 後方が一致する部分(サフィックス), および中間が一致する部分がある．EC サイトでは商品情報と商品情報の間の文字列が固定の例は見られなかったため，プレフィックスおよびサフィックスのみを抽出とする．

3.1.2 属性ごとの商品属性値の抽出箇所判定

ある商品説明ページの全体固定部分ではないタグパスにスロットが含まれると推測される．そこから取得される値の特徴およびその周辺固定部分の特徴をもとに，商品属性値が抽出されるスロットが含まれるか否かを判定する．

スロットであってもバナー広告や「ようこそ○○さん」のようにユーザごとに表示が切り替わる部分など商品情報以外も含まれる．また，商品属性であったとしてもどの属性にあたるか不明である．ただし，スロットの周辺に固定的に表れる文字列がスロットに含まれる情報が何であるかを示していたり，またスロットに入る情報の文字列的な特徴がその情報が何であるかを示していたりする場合がある．

本研究では，抽出しようとする商品属性が定まっているため，あらかじめ商品属性ごとに，周辺に現れそうな文字列や，取得される文字列の特徴を推測しておくことが可能である．そこで，本稿ではこのような商品属性ごとの特徴を手掛かりとして与えることとし，各手掛かりとスロットの周辺固定文字列やスロットから取得される文字列を比較することで，そのスロットから取得される値が商品属性の値であるかを判定する．

商品属性判定の手掛かり

手掛かりには，先に述べたように周辺固定文字列に特徴があるもの，取得される文字列に特徴があるもの，など複数が考えられる．また，多くの商品属性において有効であるものと，ある商品属性でのみ有効なものがある．

多くの商品属性で利用できる手掛かりには次のようなものがある．

＜一度だけ現れるタグパス＞

ある商品説明ページに一度だけ現れるタグパスは，特別な意味を持っていることが多い．例えば，HTML における TITLE 要素は通常 1 ページに 1 つだけ書かれ，その内容はページの主題を表す．今回対象としている商品説明ページは 1 ページで一つの商品のみを掲載するため，商品に関する情報は，ページ内で一度だけ現れるタグパスに記載されることが多い．

＜周辺文字列＞

スロット周辺にその情報の商品属性名が推測できる文字列が併記されている場合がある. この時, その文字列はテンプレートの固定部分として抽出されているはずである．具体的な文字列については，商品属性ごとに準備する必要がある．

表 3-1 にそれぞれの商品属性の毎に特徴を整理する．

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

商品名	商品に関する記述の初めの方に出現する． title 要素の内容に含まれている場合が多い． h1 要素に記述されることが多い． 属性名が周辺に現れることは少ない．
商品 コード	「コード」,「品番」等がスロット周辺に現れる． 英数字の組合せなど文字種が特徴的である．
価格	周辺に「価格」や「プライス」，または通貨記号が現れる． 文字種が特徴的である． HTML タグの class 属性値，id 属性値が文字列「price」を含む
画像	IMG 要素である． ページ内で比較的大きい画像がある場合，それが商品画像である可能性が高い．
説明文	比較的長いテキストノードである． 句読点を含む．
パンくず リスト	複数のノードから成る属性値であると考えられる． リンクアンカーとセパレータとなるテキストノードが交互に現れる特徴的な構造をもつ．

表 3-1. 商品属性ごとの特徴

3.1.3 スコアリング

前節で述べたように手掛かりは複数あり，与えられた商品説明ページによって構造が異なる場合もあることから，手掛かりが当てはまるタグパスが複数存在する可能性がある．さらに，各手掛かりが，商品属性の特徴をどれくらい示すかは異なると考えられる．例えば，あるタグパスから得られる文字列が価格の商品属性値であるかを判定する場合に，数値列であるという特徴よりも，周辺に「価格」という文字列や通貨記号が出現するという特徴の方が信頼できる．
このため，手掛かりごとにその手掛かりの信頼度を重みとして与え，図 3-2 に示す方法で商品属性ごとにある抽出箇所候補から得られる文字列が商品属性値らしいかをスコアリングし，スコアが上位のものをその商品属性の抽出箇所とする．

3.2 商品情報抽出ルール

前節までに得られた商品属性ごとに定義される抽出箇所を，まとめ商品情報抽出ルールとする．商品情報抽出ルールの表現には，XPath を使用することとする．XPath の記述方法により，商品情報抽出ルールを生成した EC サイトにデザインの変更が起きた場合や，似通ったテンプレートを利用していると想定される Web ページが存在した場合に，対応できるかが異なる場合がある．このため変更にロバストな XPath の記述について検討を行った．

1) 各 Web ページで
1-1) 手掛かりにあてはまるタグパスと文字列のペアに，手掛かりの信頼度に基づき属性値らしさの点数をつける．
1-2) 手掛かりをもとに抽出箇所の候補を作成する．
2) 全ページの抽出箇所候補を集める．
3) 各ページにから抽出箇所候補を適用し，取れたタグパスと文字列のペアの点数の合計を抽出箇所候補の点数とする

図 3-2 抽出箇所のスコアリング

3.2.1 始点の工夫

ルートを始点とする素朴な商品情報抽出ルールを作ると，ルートに近いノードに親ノードが追加・変更された場合に以下のような変更が起きたときに商品情報が全く抽出できなくなる．そこで，始点をルートから遠くとり，XPath を短く書くようにする．
HTML における id 属性値はページごとにユニークであるよう求められている．この規則は多くの Web ページで守られているので，id 属性値をもつノードを始点として，XPath を短縮することができる．
ページ横断的に統一的な利用が成されていない場合があり，商品情報を表すブロックが，商品 ID を id 属性値にもつといった不要な指定が行われていることもあるが，このような場合でも，商品の名前，価格など，ページ横断的に統一的なスタイルの指定が必要な部分には，他の id 属性値が割り当てられていることが多いので，商品 ID を id 属性値にもつノードを始点とした短縮を行ってしまうことは少なく，実用上問題はないと考えられる．

3.2.2 ポジション指定の不使用

ポジション指定とは，ノードセットの中で何番目という形のノードの指定である．例えば，次の例で /td[10] とは，td 要素のノードセットのうち「10 番目」という指定である．
/html/body/table[@class="attrs"]/tbody/tr/td[10]/a/text()
このような記述を使用すると，例えば，商品情報抽出に書かれた価格と同じノードに，セール価格が追加された場合など，何番目であるかがずれ機能しなくなる可能性がある．このため，ポジション指定をできるだけ使わないことを検討する．

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

例えば，前記の例で， /html/body/table[@class="attrs"]/tbody/tr 以下に /td/a/text() が 1 つしかない場合には，ポジション指定を省き
/html/body/table[@class="attrs"]/tbody/tr/td/a/text() と書く．

3.2.3 テキストノードの内容の利用

テンプレートは個別の商品説明ページによって変化せず，テンプレート固定部分と属性値の抽出箇所は相対的な位置は一定であると考えられる．そこで，まず固定部分を指定し，そこから抽出箇所への相対パスを記述することが有効であることが Myllymaki らの研究によりわかっている[1]．初めに指定する部分をアンカー，そこからの相対パスをホップと呼ぶ．Myllymaki らはアンカー指定に単純に文字列のみを利用しているためアンカーとして指定した文字列が，複数個所に現れると精度が落ちる．そこで，本稿ではアンカー指定にはタグパスと文字列のペアを利用することにする．ホップ指定は一度共通の祖先まで遡って，そこから抽出箇所への XPath を生成する．

4. 実験

本章ではこれまでに述べた提案手法を実装し，実際の商品説明ページを用いて商品情報の抽出ルールを自動生成し，生成された商品情報抽出ルールを利用して商品情報を抽出した結果について報告する．

4.1 データセット

65 の EC サイトごとに商品説明ページのみを 30 ページずつ収集した．収集の際のページの種類判定はその Web ページの URL や Web ページに含まれる情報の特徴を利用して自動で行っており誤りが含まれる場合がある．
20 ページを商品情報抽出ルールの自動生成用とし，残り 10 ページを評価用として自動生成した商品情報抽出ルールを適用して商品情報を抽出した．
正解データは，手動で商品情報の抽出ルールを作成しこの商品情報抽出ルールを利用して評価用データから商品情報を抽出した．

4.2 評価

次のような手順で評価を行った．

1) 提案手法を実装したプログラムに同じ種類の商品ページを入力として与え，ルールの自動生成する．
2) 評価用の Web ページに，自動生成した商品情報の抽出ルールを適用して商品情報を抽出する．
3) プログラムにより抽出された商品情報が正解であったかを判定する

	商品名	商品コード	価格	商品説明文	商品画像 URL	パンくずリスト
上位 1 ルール	0.645	0.635	0.672	0.552	0.599	0.458
上位 5 ルール	0.918	0.665	0.793	0.824	0.748	0.791

表 2 商品情報抽出結果精度

商品情報の抽出ルールは，上位 1 ルールと，上位 5 ルールの 2 種を作成した．
上位 1 ルールとは，3.1.3 項で説明したスコアリングの結果上位 1 位の商品属性値抽出箇所候補をのみを商品情報抽出ルールとして採用したものであり，上位 5 位ルールとは，上位 5 位までを商品情報抽出ルールとして採用したものである．
プログラムにより抽出された商品情報が正解であったかは，正解データと一致するかにより判定する．ただし，上位 5 位ルールによる抽出結果が正解であったかの判定については採用された 5 つのルールによって抽出された結果のいずれかと，正解データが一致すれば正解したとしている．

4.3 考察

商品名ついては上位 5 位ルールで正解データをほぼ抽出できており．商品説明文についても 8 割の精度が得られえた．
各商品属性で，上位 1 ルールによる抽出結果と，上位 5 ルールによる抽出結果を見ると，商品名，商品説明文，パンくずリストで 30%前後，価格と商品画像でも 10%以上上位 5 位ルールの結果のほうが良くなっている．手掛かりの信頼度をチューニングすることで精度を向上できるのではないかと考えられる．
商品コードは，上位 5 ルールを使用しても 67%程度の精度しか得られていない．手掛かりとして抽出値が英数字であることを与えているが，実際には英数字以外を含む商品コードが存在し，抽出できなかったためである．パンくずリストは，構造を手掛かりとして抽出を試みたが，手掛かりのバリエーションが十分ではなく，精度を得られなかった．
マーケティング分析のデータとして用いる場合，商品の価格は重要な情報となると想定されるが上位 5 位ルールでも 80%弱の精度となっている．失敗した例を見てみると，複数の価格情報が書かれている場合に，販売価格以外を取得してしまう例が多かった．EC サイトの商品説明ページにおいては，価格に関係する情報として希望小売価格，販売価格，税込価格，税抜き価格値下げ額などが複数書かれている場合が多い．この

Vol.2012-IFAT-105 No.13
Vol.2012-NL-205 No.13
2012/1/21

場合には相互の数値の関係を用いて販売価格を推定することが可能であると考えられ，商品情報抽出ルールでは複数の値を抽出し，後段の処理でその組み合わせから商品属性値を算出するような方法も検討できる．

5. おわりに

EC サイトの商品説明ページから，商品名や価格情報，商品説明文等の商品属性からなる商品情報を自動で抽出するために，商品説明ページの集合を入力として与えることで，Web ラッパーつまり商品情報抽出ルールを自動生成する方法について提案した．また，提案手法を実装し，有効性を示した．
今後は，対象とする EC サイト，商品属性を拡充した場合にも精度が維持できるかを確認し，また抽出精度を向上させるために手掛かりの拡充の検討やパラメータの調整を行う予定である．
また，今回の課題とはしなかったが，ルール作成の完全自動化のためには，商品説明ページの自動判定も必要となる．同じ種類のページであれば固定部分が共通することがわかっているので，この性質を利用した判定が可能か等の検討していきたい．

参考文献

1) N. Kushmerick: Wrapper Induction: Efficiency and Expressiveness, Artificial Intelligence, Vol.118, pp.15-68, 2000.
2) J. Myllymaki and J. Jackson. Robust web data extraction with XML path expressions. Technical Report RJ 10245, IBM, 2002.
3) R.Baumgartner, S.Flesca and G.Gottlob, Visual Web Information Extraction with Lixto, Proc. Of the 27th International Conference on Very Large Data Bases, THE VLDB Journal 2001, pp.119-128, 2001
4) S.N.Minton, S.I. Ticrea and J.Beach, Trainability: Developing a responsive learning system, Proc. Of the 18th International Conference on Artificial Intelligence 2003 Workshop on Information Integration on the Web, pp.27-32, 2003
5) http://www.w3.org/TR/2007/REC-xpath20-20070123/