HTML 文書からの商品情報抽出方式の提案

2Y－3

冨田一郎† 手塚祐一† 山本修一郎‡ 長岡満夫†

†NTT ソフトウエア研究所 ‡NTT マルチメディアシステム総合研究所

1. はじめに

本稿では、テンプレートを用いた HTML 文書からの商品情報抽出方式について述べ、抽出に用いるテンプレートを自動生成する方式を提案する。さらに、この 2 つの方式を用いた商品検索システム RBIMD (Repository Based Internet Mall Directory) について述べる。RBIMD は、WWW 上で現在運用中の複数のオンラインショップを情報源とした、商品属性のキーワード投入による商品検索サービスを提供するシステムである。

2. 商品検索の現状

WWW 上で購入希望の商品を探し出すには、その商品を取り扱うオンラインショップを探し出す必要がある。しかし、NTT DIRECTORY[1]等の WWW 上の検索サービスだけでは、欲しい情報とは全く関係ない WWW ページが数多く提示されるため、ショップを特定するのが難しい。さらに、これらの手段で知り得た複数のショップのうち、どれが希望商品を取り扱っているかを知るには、ショップを１つ１つアクセスして商品の有無を検索する必要がある。この手間に加えて、ショップが備える検索機能のインターフェイスはショップ毎に異なっているため、それぞれ把握する煩わしさが付きまとう。このように、WWW 上で欲しい商品を検索するには多大な時間と労力が必要となってしまう。

この商品検索の問題は、従来の検索サービスが、オンラインショップを含めた WWW ページ全体を対象としていることに加え、「失楽園」「渡辺淳一」といった商品情報と「書名」「著者名」といった商品属性との関係を一切考慮していないことに大きな原因がある。

3. 商品情報の抽出

我々は、オンラインショップの HTML 文書について商品情報の記述の特徴を調査し、その結果から推測される「テンプレートによる商品情報抽出方式」について抽出の確かさを調べる実験を行った。

3.1 商品情報記述の特徴

オンラインショップを調査した結果、ショップ上の商品検索機能を操作して得られる HTML 文書には、次の性質を見出すことができた。

(1) 一つのショップについて、商品検索結果の HTML 文書はすべて同じスタイルである。文書の先頭部分・末尾部分は全文書に共通で、商品記述部分のみが異なる。

(2) (1)に加え、同一ショップの商品検索結果の HTML 文書中に出現する商品記述はすべて一定のパターンで記述されており、具体的な属性値は異なるが、属性値の並び順や HTML タグの用い方が一定である。

上記の、商品検索結果の HTML 文書が持つ性質を「商品情報記述パターンの一定性」と呼ぶことにする。なお、この性質は、2 つ以上のショップにまたがって成り立つとは限らない。

3.2 商品情報の抽出実験

商品情報記述パターンの一定性より、ショップ毎に１つのテンプレートを用意すれば商品情報の抽出が可能だと考え、抽出実験を行った。

抽出対象の HTML 文書は、キーワード投入による商品検索機能を備えたオンラインショップ 13 店から商品検索を行って得たものである。オンラインショップの使用言語は日本語あるいは英語である。サンプル文書数はショップあたり１～４である。抽出用テンプレートは、図１に示すような正規表現で記述し、ショップ毎に用意した。商品情報の抽出は、HTML 文書とテンプレートである正規表現をマッチングさせ、(.*?)に対応した文字列を商品情報とした。この抽出実験の結果を表１に示す。

```
<tr><td align=center>¥s*?¥d*?</td><td>
<a href="(.*?)">(.*?)</a></td><td>(.*?)</td>
<td>(.*?)</td><td align=right>(.*?)</td>
<td>(.*?)</td></tr>
```

図 1　正規表現で記述したテンプレートの例

The Method of Product Information Extraction from HTML Documents
Ichiro Tomita, Yuichi Tezuka, Shuichiro Yamamoto, and Mitsuo Nagaoka
NTT Software Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo 180 Japan

表1 テンプレートによる商品情報抽出実験の結果

			商品数*3	抽出数*4	抽出率*5
英語	書籍	A	410	408	99.5
		B	232	223	96.1
		C	384	384	100
		D	282	272	96.5
	音*1	A	71	65	91.5
	ソ*2	A	207	207	100
		B	62	51	82.3
日本語	書籍	A	39	39	100
		B	10	10	100
		C	42	42	100
	音*1	A	9	9	100
		B	10	10	100
	ソ*2	A	10	10	100
合計			1768	1730	97.9

*1, *2: 音、ソは音楽CD、コンピュータソフトウェアを表す.
*3: HTML文書内に現れている全商品数 (個)
*4: 抽出によって情報が正しく抽出できた商品数 (個)
一つの属性値も抽出できなかった商品、および属性にそぐわない値が抽出された商品は、正しくないとした。
*5: 抽出数／商品数×100 (%)

表1の実験結果より、商品情報を正しく抽出できた商品数の割合は、大半のショップで90%以上、全体でも97.9%と非常に高かった。また、我々は、この結果より、調査したショップについては、正規表現で記述したテンプレートによる商品情報抽出方式が十分確実なものであると判断する。

4. テンプレート自動生成方式の提案

ここで、ショップ毎にテンプレートを作成する手間を省くために、HTML文書から差分および最長共通部分文字列を利用したテンプレート自動生成方式を提案する。

比較的多くのショップに当てはまるという理由より、このテンプレート生成方式は、1文書上の全商品情報が <TABLE> ~ </TABLE>, <DL> ~ </DL> 等の構文構造に囲まれているHTML文書を対象としている。なお、ここでいうHTML文書とは、ショップ側の商品検索機能を操作して得られるHTML文書である。以下に手順の概略を示す。

(1) [商品情報を含む行の特定] 同一ショップの複数のHTML文書を行単位で比較して差分を求め、互いに異なる行に印をつける。
(2) [商品情報記述全体の特定] HTML文書全体を構文解析し、(1)で得た互いに異なる行を最も多く含む構文構造を求める。
(3) [商品情報記述の商品単位への分解] (2)で求めた商品情報記述全体を構文構造に応じて商品単位に分解する。例えば、構文構造が<DL>~</DL>ならば、<DT>を目印に分解する。
(4) [各商品記述に共通な部分の特定] 各商品の記述を文字単位で比較して共通部分文字列を求める。これにより各商品情報記述に共通なHTMLタグ等の商品情報以外の文字列が判明する。
(5) [正規表現の生成] (4)で求めた最長共通部分文字列と各商品の記述との差分が抽出すべき商品情報となる。よって、最小共通部分文字列と商品記述を比較し、同一部分はそのまま、異なる部分は(.*?)で置き換える。

5. 商品情報検索システム RBIMD

我々は、これまでに述べた自動生成されるテンプレートにより商品情報抽出を行う方式を用いて、図2に示した商品情報検索システムRBIMDを開発した。

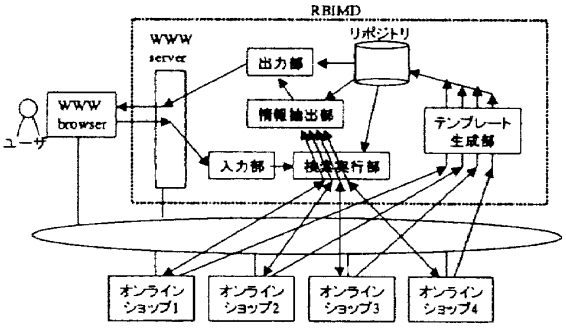

図2 商品情報検索システムRBIMD

抽出用テンプレートは、テンプレート生成部であらかじめ生成し、リポジトリ・データベース内に格納しておく。商品検索時には、ユーザから入力された商品属性のキーワードおよびリポジトリ内の各ショップに対する検索手順を用いて複数のオンラインショップ上での商品検索を行ない、得られたHTML文書からテンプレートを用いて商品情報を抽出し、整形してユーザに出力する。このRBIMDにより、ユーザは、ショップのURLやインターフェイスを意識せずに、複数ショップの商品を検索することができる。

6. まとめ

本稿では、まず、テンプレートによるHTML文書からの商品情報抽出方式について述べ、それが十分確実な抽出方法であることを実験により示した。さらに、ショップ毎にテンプレートを作成する手間を省くため、最長共通部分文字列による抽出用テンプレート自動生成方式を提案した。そして、上記2方式を用いた商品情報検索システムRBIMDを紹介した。今後は、テンプレート自動生成方式の評価およびRBIMDの評価を行う予定である。

参考文献

[1] NTT DIRECTORY, http://navi.ntt.co.jp
[2] Robert B. Doorenbos, Oren Etzioni, and Daniel S. Weld, "A Scalable Comparison-Shopping Agent for the World-Wide Web", Autonomous Agents 97