Title	カテゴリに基づく製品情報の組織化
Author(s)	有賀，忠徳
Citation	
Issue Date	2000-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1353
Rights	
Description	佐藤理史，情報科学研究科，修士

修 士 論 文

カテゴリに基づく製品情報の組織化

指導教官　佐藤理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

有賀忠徳

2000 年 3 月

要 旨

現在、WWW には多くの製品情報が存在するが、その多くは各企業のサイトに掲載されており、複数の企業の製品を比較するような用途には適していない。本稿では、それぞれの企業のサイトから製品情報が掲載されているページを収集し、それらをカテゴリ別に組織化し直し、製品一覧表を自動生成するシステムについて述べる。これを実現する中心技術は、カテゴリ別製品情報の自動収集と自動抽出である。製品情報の自動収集では、企業のページに存在するカテゴリ一覧表を利用し、製品カテゴリ別に製品ページを収集する。収集された製品ページから、テーブル解析とフォーマット解析により製品情報を抽出し、データベースに登録する。このように作成されたデータベースから、製品カテゴリ毎に組織化した製品一覧表を生成する。

目 次

第 1 章

序論

インターネットの普及により、ワールドワイドウェブ（World Wide Web; 以下 WWW と略記する）上には多くの製品情報が公開されるようになった。これらの製品情報は、消費者にとって、買物における意思決定の際に大変有用な情報となる。

買物おける意思決定には、製品選択とショップ選択の 2 つの段階が存在する。

第一段階の製品選択は、「どの製品を買うか」を決定する段階である。我々が買物を始める際、例えば「テレビを買おう」のように、既にどのような種類の製品を購入するのか決めているのが普通である。このような製品の種別を、以下では**製品カテゴリ**と呼ぶ。例えば、家電製品ではテレビやビデオなどが、情報機器ではデスクトップコンピュータやプリンタなどが、この製品カテゴリとなる。第一段階の製品選択とは、「ある製品カテゴリにおいて、購入する製品を具体的に選び出すこと」と捉えることができる。

購入する製品を決定した後は、第二段階のショップ選択に進む。ショップ選択は、「どこから（どこで）買うか」を決定する段階である。多くの製品は、複数のショップで購入可能であり、それぞれのショップにおいて、価格やそれに付帯するサービスが異なる。そのため、多くの人々は、いくつかのショップで価格やサービスを比較し、できるだけ有利な条件で購入できるショップを選ぶのが普通である。

以上のような買物の二つの段階において重要なことは、「比較する」という作業である。

製品選択の段階では、同じ製品カテゴリに属する複数の製品を比較することが必要である。一方、ショップ選択の段階では、複数のショップにおいて、ある特定の製品に関する情報（価格やサービス）を比較することが必要となる。これらの比較作業を簡単に行えるようにすることが、買物における意志決定支援の重要な要素となる。

しかしながら、現在の WWW は、このような比較支援の機能を、十分には提供できていない。

製品選択段階における比較では、ある製品カテゴリに属する複数の製品の情報が必要となる。しかし、それらは、通常、それぞれの製品を製造している複数のメーカのウェブサイト中に分散して存在している。このため、例えばテレビの購入を検討している場合、テレビを製造している複数のメーカのウェブサイトを閲覧し、そのそれぞれに対して、そのメーカが製造しているテレビにどんな製品があり、どのような特徴があるのかを調べなければならない。

ショップ選択の場合も同様である。オンラインショップの場合は、通常の販売店のように、実際に足を運んで価格を調べる必要はない。しかし、それぞれのオンラインショップのウェブサイトを閲覧し、購入を予定している製品の価格を調査しなければならないことには変わりがない。

このような煩雑さは、WWW 上の製品情報が、ユーザにとって望ましい形で組織化されていないことに原因がある。

ユーザにとって望ましい組織化とは、次のような組織化である。

- 製品選択の段階
  製品カテゴリ毎に、製品情報（製品の特徴）が一覧できるように組織化されていること。このような組織化は、同一製品カテゴリの複数の製品を比較することを容易にする。

- ショップ選択の段階
  特定の製品毎に、各ショップでの価格（やサービス）が一覧できるように組織化されていること。このような組織化により、その製品を最も安く買えるショップを容易に発見できる。

一方、現在の WWW においては、情報は、提供者毎に組織化されているとみなすことができる。

- 製品情報
  その製品を製造しているメーカ毎に組織化されている。

- 価格情報
  その製品を販売しているショップ毎に組織化されている。

このような組織化の不整合は、WWW 上の製品情報をユーザにとって使いやすい形態に整理し直すこと、すなわち、製品情報の再組織化によって解消できると考えられる。

ショップ選択での製品情報の再組織化については、Doorenbos[1] ら、冨田ら [2] の研究がある。Doorenbos らは複数のオンラインショップで、特定の製品について価格検索を行う

システムを開発した。このシステムを利用することにより、特定の製品の価格をオンラインショップを横断して比較することが可能となり、最も安く購入できるオンラインショップを容易に発見することができる。また、冨田らの研究は、複数のオンラインショップを情報源として、製品属性を考慮に入れた製品検索を実現することにより、ショップ選択の支援を行っている。

このように、ショップ選択を支援するための製品情報の再組織化にはいくつかの先行研究が存在するが、製品選択を支援する製品情報の再組織化は、これまでほとんど研究が行われていない。

本論文では、製品選択を支援するために各メーカサイトから製品情報を抽出し、それらの製品情報を再組織化するシステムを提案する。本システムは、カテゴリ毎に製品情報をまとめた**カテゴリ別製品一覧表**を自動作成することを通して、製品選択時においてユーザの意思決定を支援する。

本論文ではまず、2 章でカテゴリ別製品一覧表と、それを自動生成するシステムの概要について述べる。3 章と 4 章では、本システムを構成する主要な 2 つのモジュールについて述べる。5 章では、実際に製品情報の組織化する実験について述べ、本システムの有効性を検討する。最後に 6 章で結論と今後の課題について述べる。

第 2 章

カテゴリ別製品一覧表の自動生成

本章では、製品情報をカテゴリに基づいて組織化した**カテゴリ別製品一覧表**と、それを自動生成するシステムの概要について述べる。

2.1 製品選択における製品比較支援

まず、現在、WWW 上に掲載されている製品情報を利用して製品選択を行う場合、どのような作業が必要になるのかを、ここでは、テレビの新規購入を例にとって考えてみよう。どのテレビを購入するのかを決定するためには、次のことを行う必要がある。

1. テレビを製造しているメーカの URL を調べる。
   どのようなメーカがテレビを製造しているかをウェブを用いて調べるのは、それほど簡単ではない。多くの場合、我々が持つ一般常識を活用した方がうまくいく。メーカ名が分かれば、ウェブディレクトリや検索エンジンを利用して、メーカのウェブサイトの URL を見つけることは比較的容易である。

2. それぞれのメーカのウェブサイトを調べ、テレビの製品情報が掲載されるページを見つける。
   多くのメーカのウェブサイトは、かなりよく組織化されており、そのメーカが製造している製品の情報を見つけることは比較的容易である。しかし、多くのメーカは、10 を越える種類のテレビを製造しており、これらのページを全て見るのは、かなりの労力を必要とする。

3. 得られた情報を比較できるような形に整理する。
   これは、頭の中で行ってもよい。しかし、簡単な表を作成することは、多くの場合、

比較を容易にする。

上記したように、製品比較に必要な情報を WWW から得ることはそれほど難しいことではない。しかし、その収集には、かなりの時間と労力が必要である。この点において、現在の WWW は、製品選択の支援という機能をほとんど何も提供していないと言える。

では、どのようなことを行えば、製品選択の支援を行うことになるだろうか。

筆者が考えるシナリオは、以下のようなものである。

1. システムに、製品カテゴリを入力する。
2. システムは、その製品カテゴリに属する製品に関する情報を、一覧表の形で提示する。この一覧表のことを、以下では**カテゴリ別製品一覧表**と呼ぶ。カテゴリ別製品一覧表は、以下の条件を満たすものとする。
   (a) そのカテゴリの製品を網羅する（できるだけ多くの製品を含む）。
   (b) それぞれの製品に対して、その製品の主要情報を含む。
   (c) それぞれの製品に対して、より詳しい情報へのポインタ（ハイパーリンク）を含む。

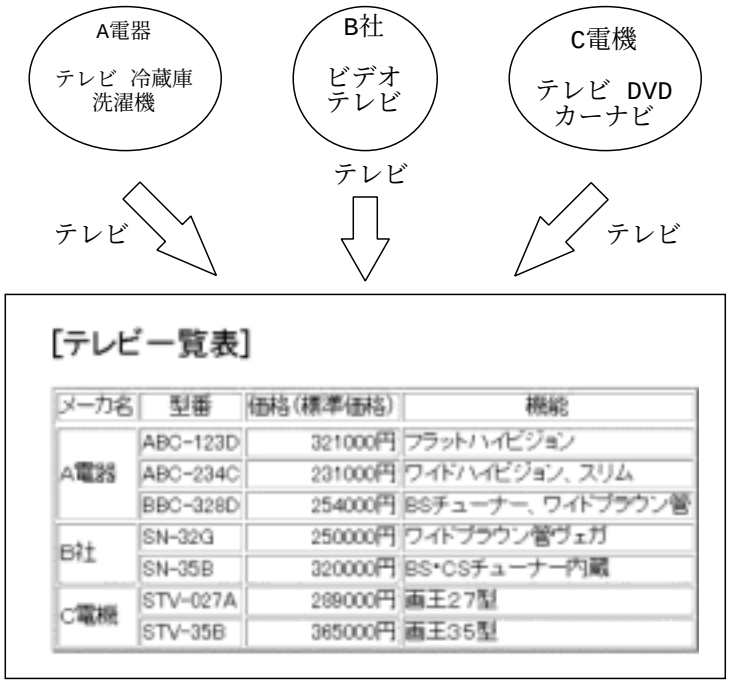
このようなカテゴリ一覧表が自動的に得られれば、製品選択は非常に容易となる。しかし、このシナリオの実現のためには、先に示した製品情報収集の作業を自動化することが必要となる。すなわち、

1. **メーカ URL の発見**
   与えられた製品カテゴリに属する製品を製造しているメーカの URL を入手する。
2. **製品ページの発見**
   それぞれのメーカのウェブサイトを調べ、その製品カテゴリの製品情報が掲載されるページを見つける。
3. **主要情報の抽出**
   それぞれのページから主要情報を抽出し、表形式に整理する。

これらの処理を行うシステムの概念図を図 2.1に示す。

2.2 システム設計

ここでは、本研究で作成したシステムの設計方針について述べる。



メーカ名	型番	価格(標準価格)	機能
A電器	ABC-123D	321000円	フラットハイビジョン
	ABC-234C	231000円	ワイドハイビジョン、スリム
	BBC-328D	254000円	BSチューナー、ワイドブラウン管
B社	SN-32G	250000円	ワイドブラウン管ヴェガ
	SN-35B	320000円	BS・CSチューナー内蔵
C電機	STV-027A	289000円	画王27型
	STV-35B	365000円	画王35型

図 2.1: システムの概念

メーカ URL の発見

前節で述べたように、ある製品カテゴリの製品を作っているメーカを、WWW を用いて調べることは、かなり難しい。そこで、本システムでは、これを自動化することをあきらめ、あらかじめシステムに与えるアプローチをとる。

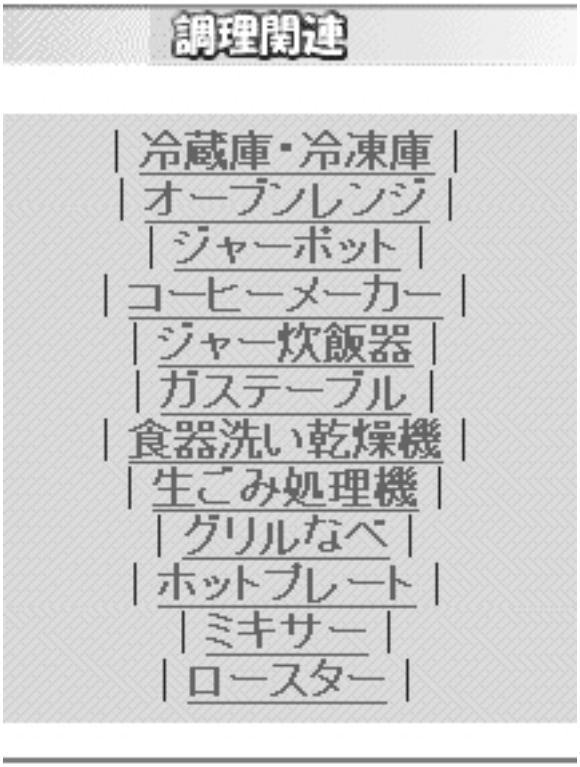
システムは、ある特定の製品カテゴリ群に対して動くように作成し、その対象領域に関する知識は、**領域知識**として、システムから独立させる。この領域知識の中に、どのようなメーカがどのような製品カテゴリの製品を製造しているかを記述しておく。また、各メーカの最上位レベルの URL（以下、**ルート URL** と呼ぶ）も、領域知識として記述しておく。

なお、これらの情報をあらかじめ領域知識として記述するアプローチをとることにより、ユーザからの入力とは独立に、あらかじめ製品情報を収集してデータベース化しておくこと（オフライン情報収集）が可能となる。本システムでは、ユーザへの応答時間短縮のため、このようなオフライン情報収集を採用する。

製品ページの発見

あるメーカのルート URL から出発し、そのメーカのウェブサイト内にある、製品情報が掲載されているページ（**製品ページ**）を発見することは、自動化する。

多くのメーカのウェブサイトは、かなり良く組織化されており、ユーザが求める情報を容易に見つけ出すことができるように工夫されているのが普通である。典型的なメーカサイトの構造を図 2.3に示す。この図に示すように、ほとんどのメーカサイトは、製造している製品の情報をカテゴリ別に整理した形で掲載しており、これらのページへのリンクを集めた**カテゴリ一覧表**が存在する。カテゴリ一覧表の例を図 2.2に示す。



http://www.hitachi.co.jp/Prod/cpim/hkjdisp5.htm

図 2.2: カテゴリ一覧表例

カテゴリ一覧表には、製品カテゴリが明示されたリンクが存在する。このため、ユーザは、この表を出発点として、比較的容易に、求める製品情報が掲載されているページへ到達することができる。このカテゴリ一覧表は、機械的に処理する場合にも有効利用できる。すなわち、カテゴリ一覧表を発見することができれば、そこからリンクを辿ることで

製品ページを発見することができる。また、それと同時に、それらの製品のカテゴリを判定することができる。

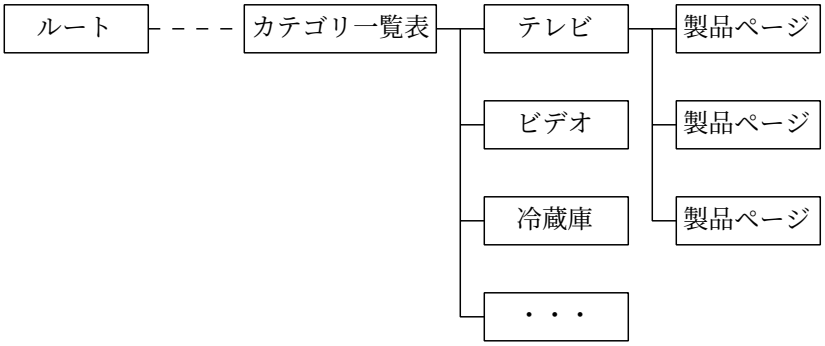


図 2.3: メーカサイトの構造

主要情報の抽出

それぞれの製品に対して、製品ページから主要な情報を抽出することも、自動化する。ここでの問題は、どのような情報を抽出するかということと、どのような方法で抽出するかということである。

我々は、複数の製品の中から一つの製品を選ぶ場合、性能、機能、デザイン、価格などの点で製品を相互に比較するのが普通である。このうち性能や機能は、それぞれの製品カテゴリにおいて、重要となる項目が異なる。例えば、テレビであれば、サイズ（画面のインチ数）は、非常に重要な機能的情報である。一方、洗濯機であれば、洗濯槽の容量が重要となる。このことから、抽出する情報の種類をそれぞれの製品カテゴリに対して定義し、その定義に従って情報を抽出することが必要となる。

本システムでは、抽出する情報を次の 2 種類に分けて定義する。

1. すべての製品カテゴリにおいて抽出する情報
   (a) 型番
   (b) 価格
   (c) 品名
2. それぞれの製品カテゴリおいて抽出する情報

一方、メーカサイトのウェブページにおいて、これらの情報は、次のような形で記述されていることが多い。

- 製品の主要な属性値
  製品の型番、価格、大きさ等の属性値は製品の特徴を知る上で大変重要な情報である。このような情報は、仕様表やページ上部の見出し（製品見出し）の中に記述されていることが多い。

- 製品を特徴づける表現
  多くの製品ページには、その製品のセールスポイントやキャッチコピー等が記述されている。これらはメーカがユーザに訴え掛けたいと考えている製品の特徴で、ユーザ側から見ると製品選択の際に有用な情報となる。このような情報は通常テキストで表現され、ページ内で強調表示されることが多い。

- 画像
  製品のデザイン。

本システムでは製品のデザインを対象外とし、製品の主要な属性値と製品を特徴づける表現の 2 種類の情報を抽出する。

2.3 システム構成

作成したシステムの構成図を図 2.4に示す。システムは、以下に示す 5 つの部分から構成される。

1. 領域知識
   対象領域（製品カテゴリ集合）に固有な情報の定義。2.4節で詳しく述べる。

2. 製品情報収集モジュール
   領域知識を利用して特定のメーカのウェブサイト内を調査し、そこに記載されている製品情報を抽出する。本モジュールは、システムの中核部分であり、以下の 2 つのサブモジュールから構成される。

   - カテゴリ URL 収集モジュール
   - 製品情報抽出モジュール

   これらのサブモジュールについては、3章と 4章において詳しく述べる。

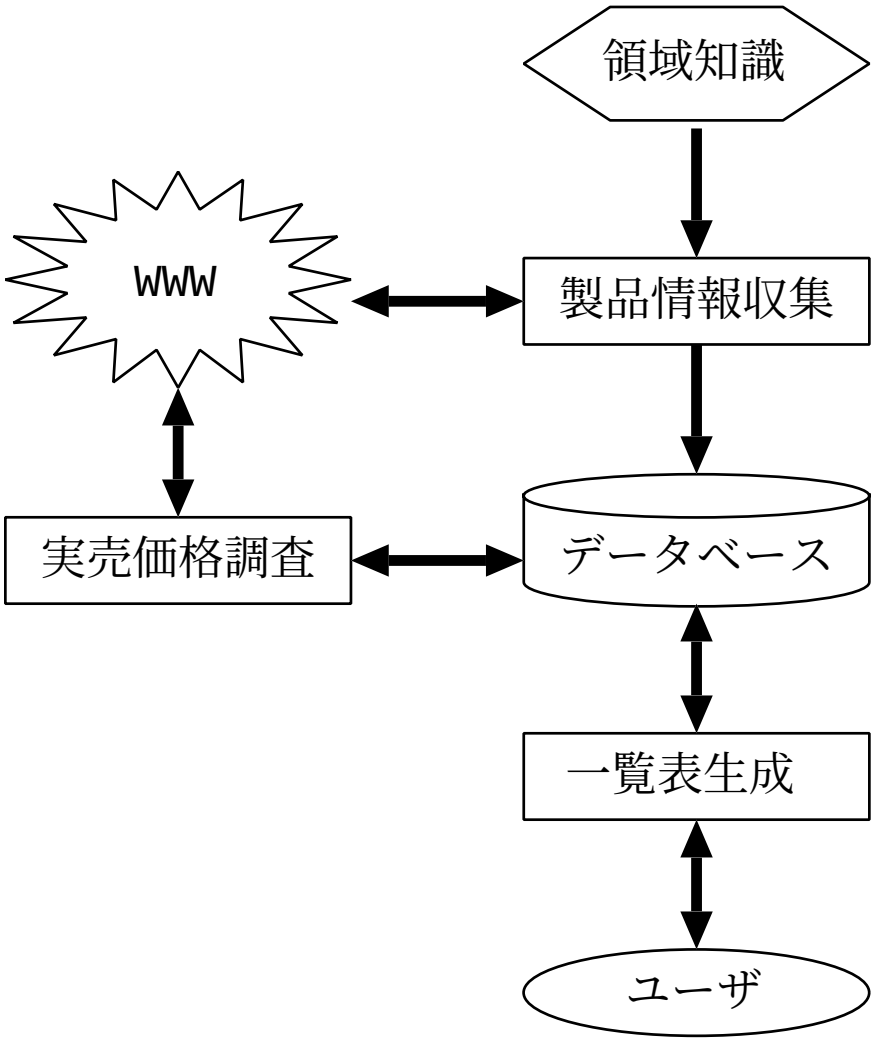


図 2.4: システム構成

3. 実売価格調査モジュール
   収集した各製品の実売価格を調査する。2.5節で詳しく述べる。

4. 製品情報データベース
   収集した製品情報と実売価格を格納する。2.6節で詳しく述べる。

5. 一覧表生成モジュール
   ユーザの要求に従ってデータベースから該当する製品を検索し、検索された製品情報を製品一覧表の形で表示する。2.7節で詳しく述べる。

本システムに対する検索要求入力ページを図 2.5に示す。このページにおいて、ユーザは、検索したい製品カテゴリ名、検索条件、ソート方法を指定する。

このページで、カテゴリ名として「テレビ」を指定した場合の出力ページを、図 2.6に示す。この図に示すように、各製品の主要な情報が表の形で整形されたものが表示されるため、ユーザは比較的容易に大まかな製品比較を行うことができる。

さらにこのページにおいて、型番は製品ページへのリンク、実売価格はオンラインショップへのリンクとなっている。このため、より詳しい製品情報が知りたければ、そのリンクをクリックすることにより、その製品情報が掲載されているページを即座に表示することができる。

2.4 領域知識

本システムは、ある特定の製品カテゴリ集合に対して動作するように設計されている。この製品カテゴリ集合のことを対象領域と呼ぶ。システムをできるだけ汎用的な形とするために、対象領域に固有な情報は、領域知識としてシステムに与えるアプローチを採る。

領域知識としてシステムに与えるものを以下に示す。

1. メーカに関する情報
   対象領域の製品を製造しているメーカに関する情報。メーカ名、メーカのルートURL、メーカの別名を記述する。記述例を以下に示す。

```
<maker>
    <name>ABC</name>
    <url>http://www.abc.com</url>
    <alias><li>BCD <li>CDE </alias>
```


Final Goal image(input) - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://tokiwa.jaist.ac.jp/~t_aruga/cgi-bin/HomeElectro2.cgi
移動
Project Home Electronics
検索する製品
検索開始
検索条件を指定
企業選択
すべて SONY HITACHI SANYO TOSHIBA
PANASONIC SHARP VICTOR
表示する項目
すべて 機能 解説 実売価格・リンク
ソート方法
メーカごと
ページが表示されました
インターネット

図 2.5: 入力画面


検索結果 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) C:¥document¥My Documents¥test.html
PROJECT WIT Home Electronics
検索結果
検索キーワード『テレビ』検索結果は 1 件です
テレビ
テレビの結果
Maker
型番
価格
品名
Product_Caption
売値
SHARP
32C-PB1
230,000
フラットワイドテレビ
BSデジタル放送対応次世代テレビ「DOMANI(ドマーニ)」
シリーズ 3機種を一斉新発売
[高画質]
149,800
175,000
SHARP
28C-WZ1
165,000
ワイドテレビ
28型,32型 BS内蔵ワイドテレビ2機種を発売
画面の明るさを抑え,消費電力を最大15W抑える「明るさひかえめボタン」搭載。
瞬時に最適なワイド映像に切り替える「ワイドクリアビジョン放送識別回路」内蔵。
109,800
119,800
PANASONIC
TH-47FP10
560,000
大画面テレビ ハイビジョンプロジェクション
-
VICTOR
HV-D50LA1
850,000
LAプロジェクション テレビ
大画面時代に対応する業界最高の画質を実現した 新開発"D-ILA ホログラム素子"搭載
見る場所を選ばない業界最高クラスの広視野角を実現ゴースト・リダクション・チューナー
514,800
SONY
SU-28F1
オープン
FDトリニトロン管搭載ワイドテレビ KW-28HDF7
176,800
SONY
KV-25DW1
115,000
ワイドテレビ ワイドトリニトロンカラーテレビ
78,000
HITACHI
W28-GF3X
175,000
2000年デジタルを先取り！ 新世紀の高画質テレビ。
BSデジタル放送に高画質・高音質対応！
現行放送も目にやさしいプログレッシブ高画質に
109,800
115,800
2000年デジタルを先取り！ 新世紀の高画質テレビ。

図 2.6: 結果画面

</maker>

この例は、メーカ名が「ABC」、ルート URL が「http://www.abc.com」、メーカの別名が「BCD」、「CDE」であることを表現している。

2. 製品カテゴリに関する情報
対象領域の製品を分類するカテゴリに関する情報。カテゴリ名とそのカテゴリ固有の特徴フィールドを記述する。記述例を以下に示す。

<category>
<name> 洗濯機 </name>
<field> 容量 <li> 大きさ </field>
</category>

この例は、対象領域のなかに「洗濯機」というカテゴリが存在し、その特徴フィールドとて「容量」と「大きさ」を持つことを表現している。

3. テーブルヘッダに関する情報
製品情報の抽出元となるテーブルで用いられるヘッダに関する情報。テーブルヘッダ名、テーブル抽出に用いる正規表現、テーブルヘッダの単位を記述する。記述例を以下に示す。

<table_header>
<name> 価格 </name>
<pattern>(?:^|[^\dA-Z])([\dA-Z]{0,3}-[\dA-Z]{1,6})(?:$|[^\dA-Z])
</pattern>
</table_header>

この例は、「価格」というテーブルヘッダが存在し、その値を抽出するための正規表現が

```
(?:^|[^\dA-Z])([\dA-Z]{0,3}-[\dA-Z]{1,6})(?:$|[^\dA-Z])
```

であることを表現している。

4. 共通フィールドに関する情報。
本領域で共通に用いるフィールドに関する情報を記述する。記述例を以下に示す。

<field key=true><header_name> 型番 </header_name></field>

この例は、「型番」が共通フィールドとして使用されることを表現している。また、例のように、属性値に「key=true」が設定されているフィールドが製品識別子として用いられる。

現在のシステムで用いている領域知識のすべてを付録 Aに示す。

2.5 実売価格調査

実売価格調査モジュールは、製品情報収集モジュールによって抽出された各々の製品の実売価格を調査するモジュールである。

メーカサイトから得られる製品情報の中には、定価が含まれることもあるが、実売価格は一般に定価と異なるのが普通である。また、近年は、定価を定めない（オープン価格の）製品も多い。このため、実売価格は、メーカサイトから得られない。しかし、実売価格は、どの製品を購入するのかを決定する際の最も重要な情報の一つであるため、本システムにおいてこれを提供することは不可欠である。

ウェブ上には、多くのオンラインショップが存在し、各社の製品を販売している。そこで、このオンラインショップを利用して、実売価格を調査する方法が考えられる。オンラインショップにおける典型的なページの例を図 2.7に示す。この図のように、オンラインショップでは、取り扱っている製品と価格を表形式で掲載している場合が多い。つまり、調べようとしている製品がこの表のどの列に対応しているのかが決定できれば、実売価格を得ることができる。

ここでは、製品の特定に、メーカ名と製品番号（型番）を利用し、以下の方法で実売価格を調べる。

1. メーカ名と製品番号から成る検索質問（クエリ）を検索エンジンに入力し、検索された URL を収集する。現在、検索エンジンとしては、infoseek[1]を使用し、最大 50 件の URL を収集する。入力する検索質問は、次の通りである。

   メーカ名 AND 製品番号

2. こうして得られた URL のソースをすべてダウンロードする。

3. それぞれのページに対して、以下の処理を行う。

[1]http://www.infoseek.co.jp

冷蔵庫 - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

アドレス(D) http://www.fifty-fifty.com/shop/reizoko/ 移動

冷蔵庫

メーカー	型 番	定 価	特別価格(税別)	備 考
HITACHI	R-5P1	¥36,000	¥23,800	46L 1ドア
HITACHI	R-8P1	¥41,000		81L 1ドア
HITACHI	R-8T3	OP	¥24,500	80L 2ドア
HITACHI	R-127A	OP	¥28,800	120L 2ドア
HITACHI	R-162A	OP	¥34,800	160L 2ドア
HITACHI	R-22YK	¥110,000	¥64,500 New	220L
HITACHI	R-41EPAM	¥200,000	¥122,500 New	410L PAM
HITACHI	R-S41EPAM	¥225,000	¥138,000 New	410L PAM 自動製氷
HITACHI	R-S47EPAM	¥260,000	¥158,800 New	470L PAM 自動製氷
National	NR-A5T2	OP	¥19,800	49L 1ドア
National	NR-A7T3	OP	¥20,500	74L 1ドア
National	NR-B8T3	OP	¥22,800	78L 2ドア
National	NR-B13T3	OP	¥31,800	126L 2ドア
National	NR-B14B2	OP	¥34,800	137L 2ドア
National	NR-B14BA	OP	¥41,000	137L 2ドア
National	NR-B22T1	OP	¥57,800	220L
National	NR-C25MA	¥120,000	¥68,800	249L
National	NR-C25T1	OP	¥63,800	250L
National	NR-E40S2	¥320,000	¥212,000	400L自動製氷
National	NR-E54M2	¥465,000	¥313,000	540L自動製氷
National	NR-E40V1	¥230,000	¥123,000	401L Tanto自動製氷
National	NR-E46V1	¥265,000	¥142,000	460L Tanto自動製氷
National	NR-C32D1	¥132,000	¥89,800 New	320L Tanto自動製氷
National	NR-E35D1	¥190,000	¥128,000 New	351L Tanto自動製氷
National	NR-D36D1	¥165,000	¥111,000 New	365L Tanto自動製氷

ページが表示されました インターネット

http://www.fifty-fifty.com/shop/reizoko/

図 2.7: オンラインショップのページ例

- そのページにある表 (table タグによって書かれているもの）を抽出する。
- 4.3.1節の仕様表解析を利用し、型番に対応する実売価格を取得する。
- 得られた実売価格とそれが掲載されていたページの URL を組にして抽出する。

4. 抽出されたすべての組をデータベースに登録する。

2.6 製品情報データベース

製品情報のオフラインの調査（製品情報収集と実売価格調査）とユーザのオンライン検索を結び付けるのが、製品情報データベースである。本システムが提供する製品情報は、すべて製品情報データベースに格納されている。

本データベースは、一つのテーブルから構成される。このテーブルの 1 レコードは対象領域に依存しない 5 つのフィールドと、領域知識で定義される領域固有フィールドから成る。対象領域に依存しない 5 つのフィールドを以下に示す。

1. メーカ名
2. カテゴリ名
3. セールスポイント
4. 製品情報抽出元 URL
5. 実売価格

領域知識で定義される領域固有フィールドは、さらに、その領域すべてにおいて共通に使用するフィールド（共通フィールド）と、その領域のそれぞれのカテゴリに依存するフィールド（特徴フィールド）の 2 種類に分けられる。

現在対象としている「家電」領域においては、共通フィールドとして次の 3 つを定義している。

- 型番
- 価格
- 品名

表 2.1: 特徴フィールド例

カテゴリ名	特徴フィールド
テレビ	大きさ
洗濯機	容量
冷蔵庫	容量
掃除機	仕事率
電子レンジ	電力 容量

Maker	Category	型番	価格	品名	特徴	セールス ポイント	URL	実売 価格
HITACHI	テレビ	W28-GF	175,000		28 型	BS デジタル...	http://...	...
SHARP	テレビ	32C-PB	230,000	フラット	32 型	DOMAN...	http://...	...
SHARP	テレビ	28C-WZ	165,000	ワイド	28 型	画面の明...	http://...	...
PANASONIC	テレビ	TH-47F	560,000		47 型	大画面...	http://...	...
VICTOR	テレビ	HV-D50	850,000	ILA	50 型	ホログラム...	http://...	...

図 2.8: 製品情報データベース

一方、特徴フィールドは、それぞれのカテゴリにおいて定義されている。その例を表に示す。特徴フィールドは、一つのカテゴリに対して最大 5 個まで定義することができる。なお、本データベースにおいては、メーカ名と型番の組合わせを識別子として用いる。これは、一つのメーカにおいて、同じ型番を持つ製品は存在しないと考えることができるためである。

本データベースの一部を図 2.8に示す。

2.7 一覧表生成

一覧表生成は、ユーザの要求に従ってデータベースから該当する製品を検索し、製品一覧表を作成して表示することを行う。

ユーザからの要求としては、検索するカテゴリ名、検索対象とするメーカ、ソート方法、の 3 つを受付ける。

これらの要求を受け取るとシステムは、カテゴリ名とメーカ名から検索質問を作る。この検索質問を用いてデータベース検索を行い、得た結果から一覧表を作成する。この時、それぞれの製品の型番は製品ページへのハイパーリンク、実売価格はオンラインショップへのハイパーリンク、を付加する。また、メーカ（デフォルト）、価格、実売価格、のいづれかでソートを行う。

第 3 章

カテゴリ URL の収集

本章では、ある特定のメーカサイトにおいて、それぞれの製品カテゴリに対応する URL を収集する処理について述べる。

3.1 概要

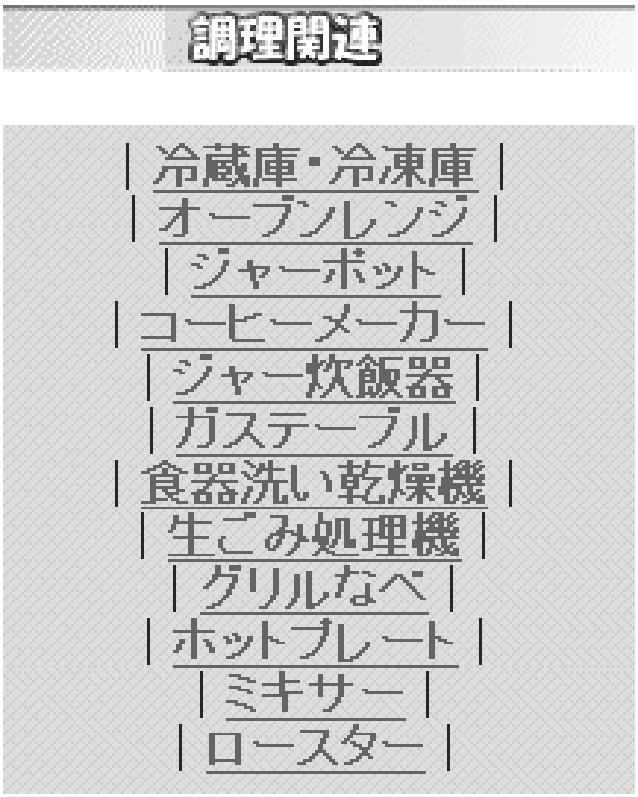
ある特定のメーカサイトから製品情報を収集するための処理の前半部は、それぞれの製品カテゴリに対応する URL（以下、**カテゴリ URL** と呼ぶ）を抽出する処理である。このカテゴリ URL は、4章で述べる製品情報抽出において、次の 2 点で重要な役割を果たす。

1. 製品情報を掲載しているページを探す出発点を与える。
2. それぞれの製品のカテゴリを決定する。（このカテゴリ URL から出発して抽出された製品情報は、そのカテゴリの製品である。）

ほとんどのメーカサイトにおいて、扱っている製品の情報はカテゴリ別に整理された形で掲載されている。そして、これらを俯瞰する目次として、**カテゴリ一覧表**が存在する。カテゴリ一覧表の例を図 3.1に示す。この図に示すように、カテゴリ一覧表は、それぞれのカテゴリに対するリンクを持つ。このリンク先ページの URL が、本章で述べる処理で収集するカテゴリ URL である。

カテゴリ URL 収集の手順の概要を図 3.2に示す。この図に示すように、手順は、以下に示す 5 つのステップからなる。

1. ページ収集
   メーカルート URL を起点としてリンクを辿り、ある条件を満たすページ群を収集する。



http://www.hitachi.co.jp/Prod/cpim/hkjdisp5.htm

図 3.1: カテゴリ一覧表例

2. 表抽出
   収集したそれぞれのページから、表を抽出する。この表が、カテゴリ一覧表の候補となる。

3. タイプ判定
   抽出した表を 4 つのタイプに分類する。

4. 一覧表判定
   抽出した表がカテゴリ一覧表かどうか判定する。

5. カテゴリ URL 抽出
   カテゴリ一覧表と判定した場合、その一覧表から、カテゴリ名とカテゴリ URL の組を抽出する。

このようにして抽出したカテゴリとカテゴリ URL の組は、４章で述べる製品情報抽出モジュールへ渡される。

以下では、まず、上記の各ステップについて、詳しく説明し、次に、カテゴリ URL 収集の実験とその結果について述べる。

3.2 ページ収集

カテゴリ URL 収集の第一ステップは、ページ収集である。ページ収集では、カテゴリ一覧表が掲載されている可能性があるページを収集する。

カテゴリ一覧表は「製品ページへのリンクの起点となる」という性質上、掲載されているページは、メーカサイトのルートページから１、２回リンクを辿ることによって到達できるような場所に存在することが多い。また、それらのリンクには、人間が見たならば、その先にカテゴリ一覧ページがあると予想できるようなラベルが付けられているのが普通である。しかし、このようなラベルの種類は予想以上に多く、また、アンカテキストとしてではなく画像として表現されている場合も多い。このため、リンクラベルを頼りに選択的にカテゴリ一覧ページを探索する方法を採用することはできない。

このような理由から、カテゴリ一覧表ページを探す基本的な探索戦略として、ある一定の深さまであらゆるリンクを辿ってページを収集し、そこで得られたページにカテゴリ一覧表が存在するかどうかを判定する方法を採用する。

但し、先に述べたリンクラベルがテキストとして得られ、かつ、その先にカテゴリ一覧ページがほぼ存在しないと思われる場合は、そのリンクを枝刈りする。また、リンクラベ

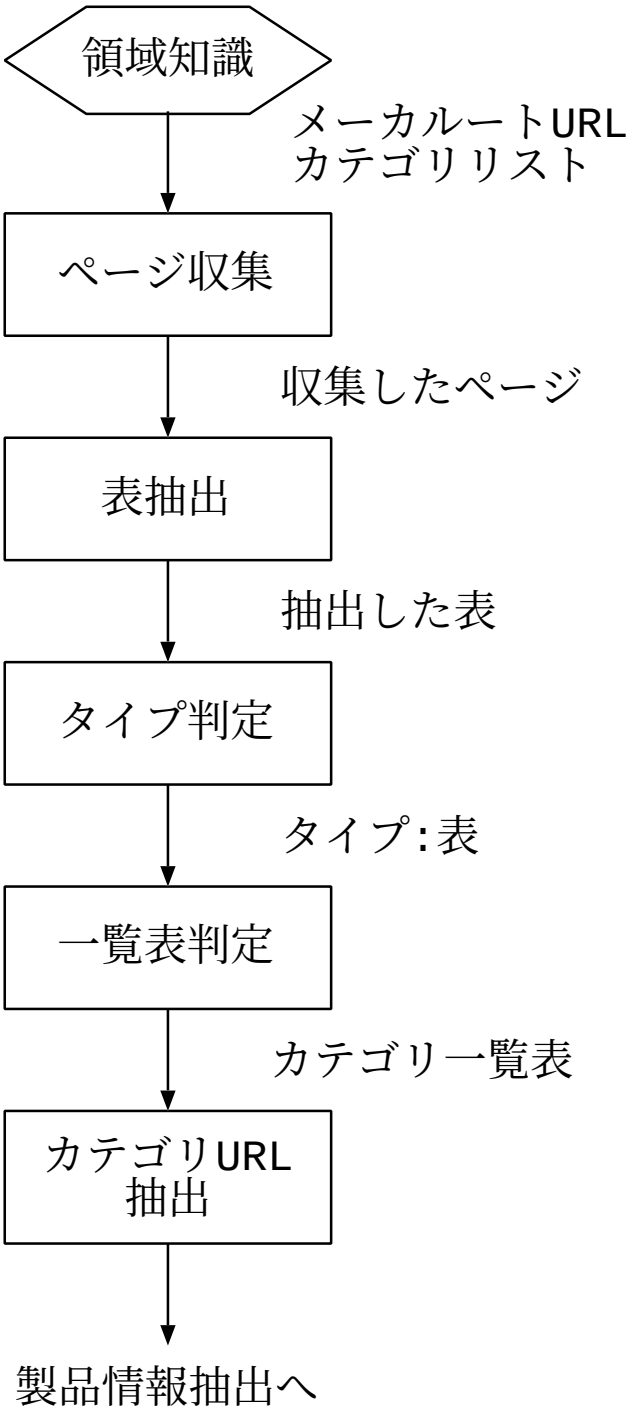


図 3.2: カテゴリ URL 収集概要

ルから、そのリンクが有望と思われる場合は、深さに関する限界を緩める操作を行う。

以下にページ収集アルゴリズムを示す。

1. URL から、そのページ（ソース）をダウンロードする。このページを候補ページに含める。

2. そのページに含まれるリンクを抽出する。抽出するリンクは、次の 3 種類である。

   (a) <a href>（a タグ）

   (b) <form> 中の <select>（select タグ）

   (c) <area>（area タグ）

3. アンカ文字列[1]が以下の正規表現にマッチした場合、そのリンクを捨てる。

   ((ホーム |HOME|TOP)(へ | ページ |PAGE))| 戻る | お問い合わせ |
   窓口 | メール | メイル | サポート | オプション | 別売 | アクセサリ)

4. アンカ文字列が以下の正規表現にマッチする場合、起点からの深さを見かけ上 1 減少させる。（このリンクは深さを増加させないリンクとして扱う。）

   製品 | 一覧 | カタログ | プロダクト |product

5. 起点 URL からの深さが深さ限界未満であれば、残ったリンクの URL に対して、同様の操作を繰り返す。

なお、家電製品を製造するメーカ 4 社のサイトに存在するカテゴリ一覧表について調査を行った結果から、深さ限界の値として 3 を用いている。

3.3 表抽出

カテゴリ URL 収集の第二ステップは、表抽出である。表抽出では、ページ収集で得られたそれぞれのページから、表を抽出する。こうして得られた表は、次のステップでそのタイプを判定し、最終的に、カテゴリ一覧表かどうか判定する。

抽出する表は、以下の 3 種類に分けられる。

[1] a タグに囲まれた文字列、select タグに囲まれた文字列、area タグ中の alt 属性で表される文字列。

1. table タグによって記述された表

2. form タグ中の select タグによって記述された表（選択肢）

3. リストタグ（ol タグ、ul タグ、dl タグ）によって記述された表（リスト）

このうち、table タグ、リストタグは表を表現するのではなく、ページのレイアウトを整えるためにも使用されることがある。これは、以下の 2 つの場合に分けられる。

1. 表のレイアウトを整えるために使用されている場合
   この場合は、table タグやリストタグが入れ子構造を形成する。これは、さらに、次の 2 つの場合に分けられる。

   (a) 外側のタグがレイアウトとして使用されている場合
   内側のタグで作成される表の配置を調整するために、外側のタグがレイアウト用として使用される。

   (b) 内側のタグがレイアウトとして使用されている場合
   表内部のデータの配置を調整するために、内側のタグがレイアウト用として使用される。

2. 表以外のレイアウトを整えるために使用されている場合
   この場合は、ページ内の table タグやリストタグは、ページレイアウト用のタグということになる。このようなタグを抽出しても、以降の一覧表判定の処理において、一覧表ではないと判定される。すなわち、表抽出のステップにおいて、このような表を抽出しても問題は発生しない。

上記の 1 の場合は、レイアウト用のタグを除去することが必要となる。これは、以下の方法で行う。

- table タグ

  1. 外側の table タグに罫線があり、かつ内側の table タグに罫線が無い場合は、内側の table タグがレイアウト用だと判断し削除する。

  2. その他の場合は、外側の table タグがレイアウト用だと判断し削除する。

- リストタグ
  外側のリストタグを削除する。

3.4 タイプ判定

カテゴリ URL 収集の第三のステップは、タイプ判定である。タイプ判定では、表抽出において抽出された表のタイプを決定する。このタイプは、以降の処理（一覧表判定とカテゴリ URL 抽出）で利用される。

ここでは、抽出した表を、次の４つのタイプに分類する。

1. タイプ 1
   table タグを使用していて、カテゴリ名がアンカ文字列に含まれる。カテゴリとリンクが 1 対 1 で対応している。例を図 3.3に示す。

2. タイプ 2
   table タグを使用していて、カテゴリ名がアンカ文字列に含まれない。一つのカテゴリ名に対して、複数のリンク先がある場合に用いられることが多い。例を図 3.4に示す。

3. タイプ 3
   select タグを使用している。カテゴリとリンクが 1 対 1 で対応している。例を図 3.5に示す。

4. タイプ 4
   リストタグを使用している。リンク先に関する簡単な説明文が掲載されていることが多い。例を図 3.6に示す。

タイプ判定は、以下のアルゴリズムで行う。

1. 表を表現しているタグを調べる。
   (a) select タグの場合は、タイプ 3 とする。
   (b) リストタグの場合は、タイプ 4 とする。
2. table タグの場合は、以下の処理を行う。
   - 表の中からアンカ文字列を抽出する
   - 抽出したアンカ文字列にカテゴリ名が含まれている場合はタイプ 1、含まれてない場合はタイプ 2 とする。

家庭電化製品

●冷蔵庫	●レンジ
●ホットプレート	●食器洗い乾燥機
●空気清浄機	●除湿器
●洗濯機	●クリーナー
●エアコン室内機・室外機	●石油暖房機
●加湿器	●ホットカーペット
●ファンヒーター	●太陽光発電システム

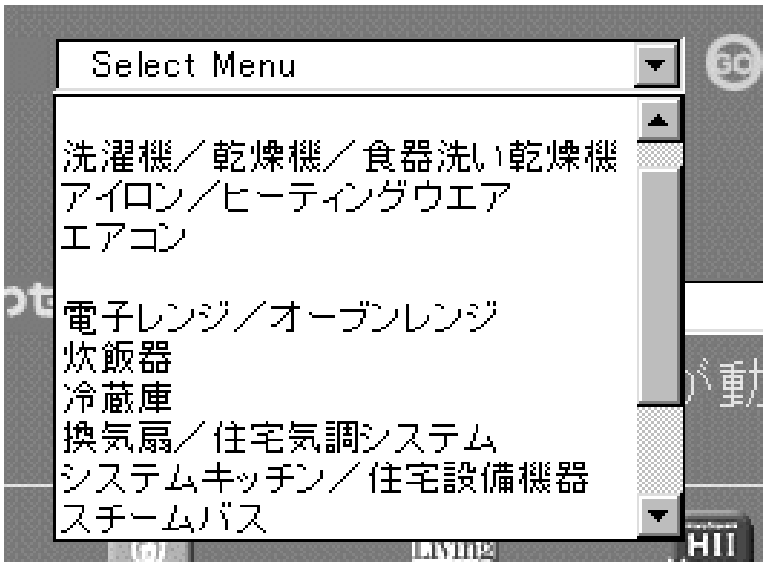
http://www.sharp.co.jp/sc/seihin/index.html

図 3.3: タイプ 1 の例

AV機器

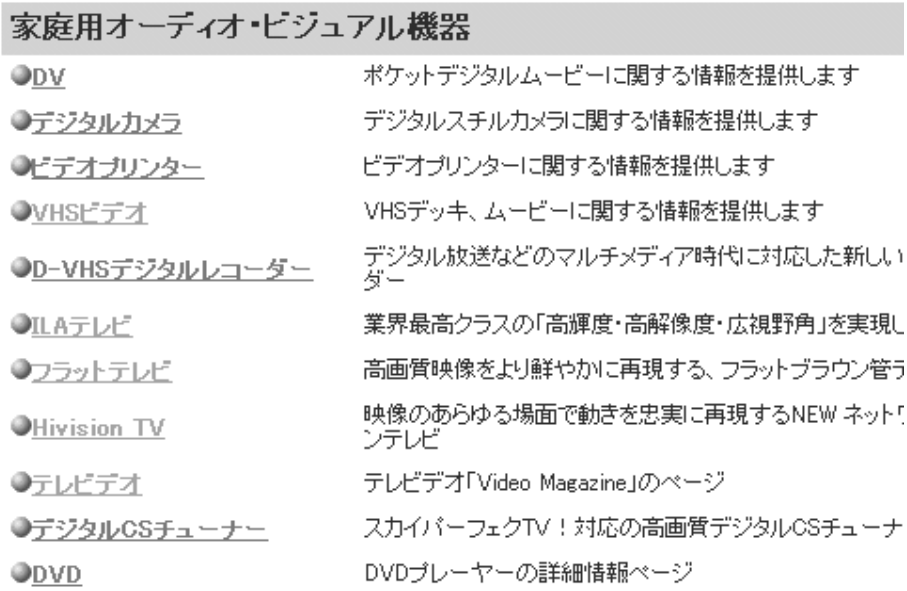
品名	新製品	ニュースリリース	特集
テレビ	○	○	○
ビデオデッキ	○	○	○
ビデオカメラ		○	
DVDプレーヤー	○	○	○
カーナビゲーションシステム		○	○
MDプレーヤー		○	
液晶ビジョン		○	

図 3.4: タイプ 2 の例



http://www.panasonic.co.jp/corp/products/main3.html

図 3.5: タイプ 3 の例



http://www.victor.co.jp/products.html

図 3.6: タイプ 4 の例

3.5 一覧表判定

カテゴリ URL 収集の第 4 のステップは、一覧表判定である。一覧表判定では、表抽出で抽出した表がカテゴリ一覧表かどうか判定する。

この判定を行う方法を決定するために、まず、家電製品を製造するメーカ 4 社のサイトに存在するカテゴリ一覧表について調査を行った。この結果、カテゴリ一覧表には、次のような表層的特徴が見られることが分かった。

特徴 1 カテゴリ名やリンクが一列（または一行）に並んでいる。

特徴 2 カテゴリ名とリンク数のバランスがとれている。

特徴 3 簡潔な表現が用いられる。

本研究では、これらの特徴を利用し、以下の方法でカテゴリ一覧表かどうかの判定を行う。

1. その表に含まれるカテゴリ数、リンク数、1 レコード当りの平均文字数、を調べる。（このカテゴリ数を調査する際に、領域知識として定義したカテゴリリストを用いる。）

2. 以下の条件をすべて満たしている場合、カテゴリ一覧表と判定する。

   条件 1 以下の条件式を満たす。

   $$\text{カテゴリ数} \times \text{倍率} > \text{リンク数}$$

   この条件式は、「リンク数に対して十分な数のカテゴリが表中に存在する」ことを意味している。カテゴリ一覧表には、少なくともカテゴリ名の数と同じ数のリンクが存在すると考えられる。しかし、領域知識として与えられるカテゴリリストは完全ではないので、表中のすべてのカテゴリ名が正しくカテゴリと判定できるとは限らない。一方、HTML で記述されているリンクは、機械的にほぼ完全に抽出することができる。この差を考慮し、抽出できたカテゴリ数を何倍かした数がリンク数を上回っている場合、リンク数に対し十分なカテゴリが表中に存在する、と判定する。

   家電製品を製造するメーカ 4 社のサイトに存在するカテゴリ一覧表について調査を行った結果を基に、倍率は一覧表のタイプによって表 3.1のように設定した。

表 3.1: 倍率

タイプ 1	3
タイプ 2	7
タイプ 3	5

条件 2 リンク数、カテゴリ数がどちらも 2 以上である。
条件 1 はカテゴリ数とリンク数の大小関係を用いているだけなので、カテゴリ数とリンク数がともに 1 である場合も条件を満たしてしまう。このような表を排除するために、最低数の制限を設ける。

条件 3 テーブルの 1 セル[2] 当りの平均文字数が 15 文字以下。
これは、カテゴリ一覧表が、「簡潔な表現が用いられる」という特徴を持つということを条件として反映させるためである。

3.6 カテゴリ URL 抽出

カテゴリ URL 収集の最後のステップは、カテゴリ URL 抽出である。カテゴリ URL 抽出では、カテゴリ一覧表と判定された表から、カテゴリ名とカテゴリ URL の組を抽出する。

抽出方法は、カテゴリ一覧表のタイプによって異なる。

1. カテゴリ一覧表のタイプが、タイプ 1、タイプ 3、タイプ 4 のいづれかの場合。
これらのタイプのカテゴリ一覧表では、カテゴリ名とリンクが一対一に対応している。そのため、HTML タグを解析することにより、簡単に、カテゴリ名とカテゴリ URL の抽出を行うことができる。

2. カテゴリ一覧表のタイプが、タイプ 2 の場合。
以下の手順で抽出を行う。

   (a) 表の方向を判定を行い、カテゴリ名が縦に並んだ表に変換する。
   3.5節で説明したように、リンクの方がカテゴリ名よりも抽出精度が高い。そこで表の方向は、列数に対するリンク数の割合、行数に対するリンク数の割合、を比較し判定する。列と行でリンク数の割合が高い方を表の向きとする。

[2] td または th タグによって囲まれた表の一項目

(b) リンク列とカテゴリ列を調査する。

リンク列、カテゴリ列とは、それぞれリンクやカテゴリ名が列記された列のことを指す。図 3.4では、左から数えて 1 列目がカテゴリ列、2 から 4 列目がリンク列である。抽出した表の各列を以下の手順で調査する。

i. リンク数が最も多い列を見つけ出し、その列をリンク列とする。（この列を最大リンク列と呼ぶ）

ii. 各列に対して、リンクに用いられる HTML タグと領域知識内のカテゴリリストを用い、リンク数とカテゴリ数を数える。

iii. 抽出したリンクの数が最大リンク列の 2 分の 1 以上の場合、その列をリンク列とする。

iv. 抽出したカテゴリ数が最大リンク列の 4 分の 1 以上の場合、その列をカテゴリ列とする。

v. 行毎にカテゴリ名とリンクを対応させる。

カテゴリ名とカテゴリ URL の抽出を行った後は、抽出したカテゴリ名、カテゴリ URL を整理する。一般的に一つの製品カテゴリには、メーカによっていくつかの異なった名称（別名）が存在する。領域知識にこれらの別名を登録しておくことによって、カテゴリ名を統一する。例えば、テレビには TV、テレビジョンといった別名が存在するが、これらをテレビに統一する。

また、メーカサイトにおいては、製品ページに到達するための経路を複数用意する傾向がある。このため、カテゴリ一覧表は一つのメーカサイトに複数存在する場合があり、あるカテゴリに対するカテゴリ URL が重複して抽出される場合がある。そこで、これらの重複を削除することを行う。

第 4 章

製品情報の抽出

本章では、3章で述べたカテゴリ URL 収集によって得られたそれぞれのカテゴリ URL を起点として、製品情報が掲載されているページを探索し、製品情報を抽出する処理について述べる。

4.1 概要

ある特定のメーカサイトから製品情報を抽出するための処理の後半部は、カテゴリ URL を起点として、そのカテゴリに属する製品の情報を抽出する処理である。

製品情報の抽出の概要を図 4.1に示す。この図に示すように、収集手順は2つのステップから構成される。

1. 製品ページ収集
   カテゴリ URL を出発点として、以下の手順で製品情報が掲載されているページ（以下、**製品ページ**と呼ぶ）を収集する。
   (a) 製品一覧表を探索する。見つかった場合は、この表より製品ページを収集する。
   (b) 製品一覧表が見つからなかった場合、ページ収集と製品ページ判定（得点付け）により、製品ページを収集する。
2. 製品情報抽出
   収集した製品ページから製品情報を抽出する。抽出処理は、次の 3 つの処理から構成される。

- 仕様表解析
  仕様表から製品の主要属性を抽出する。
- ページ見出し解析
  ページ見出しから製品の主要属性を抽出する。
- 強調文字抽出
  製品ページ内の強調された文字を抽出することにより、製品のセールスポイントを抽出する。

なお、抽出した製品情報は、メーカ名と型番を識別子として整理し、抽出元ページの URL を付加して、製品情報データベースに格納する。

4.2 製品ページ収集

製品情報抽出の第一ステップは、製品ページ収集である。製品ページ収集では、カテゴリ URL を起点として、そのカテゴリに属する製品の情報が掲載されているページを収集する。

カテゴリ URL と製品ページの関係には、次のような場合がある。

1. カテゴリ URL ページが、そのまま製品ページとなっている。
2. カテゴリ URL ページ上に存在するリンクの先に、製品ページが存在する。
3. カテゴリ URL ページ、または、そのリンク先ページに、そのカテゴリに含まれる製品を一覧するような目次（以下、**製品一覧表**と呼ぶ）が存在し、その先に、製品ページが存在する。

そのカテゴリに属する製品の種類が少ない場合は、1 である場合が多い。一方、製品の種類が多い場合は、2 をとることが多い。3 は、2 の特殊な場合で、製品一覧表という形式の目次が存在する場合である。

製品一覧表の典型的な例を図 4.2に示す。この図に示すように、製品一覧表には製品の型番と製品ページへのリンクが掲載されている。つまり、このような製品一覧表を見つけることができれば、製品の型番とその製品に関する情報を掲載したページ（製品ページ）への URL を効率よく収集することができる。

しかし、先に述べた通り、必ずしも製品一覧表が存在するとは限らない。このため、製品一覧表を発見できなかった場合は、これとは異なる方法で製品ページを収集しなければ

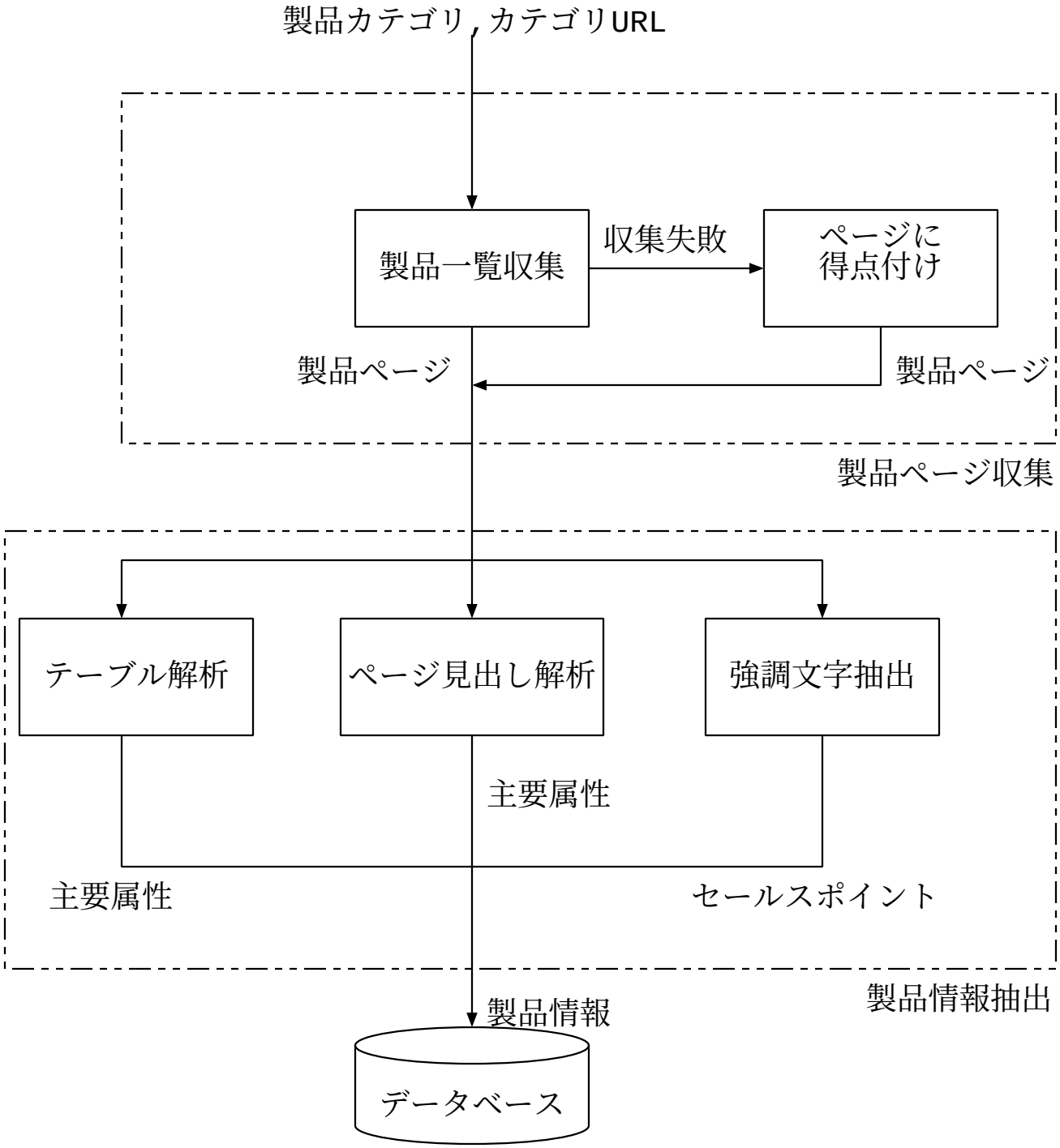


図 4.1: 製品情報抽出概要

家庭電化製品 - Microsoft Internet Explorer

ファイル(F)　編集(E)　表示(V)　お気に入り(A)　ツール(T)　ヘルプ(H)

洗濯機

品名	型 番	新製品	ニュースリリース	特集
高トルクDDインバーター全自動洗濯機	ES-SE91-C (ホワイティベージュ)		○	
	ES-SE81-C (ホワイティベージュ)		○	○
	ES-SE71-C (ホワイティベージュ)		○	
全自動一体型乾燥洗濯機	ES-E62-W-5 (ホワイト)		○	
	ES-E62-W-6 (ホワイト)		○	
全自動洗濯機	ES-A80E		○	
	ES-A70E		○	
	ES-S7PB-C (ホワイティベージュ)		○	
	ES-S6PB-C (ホワイティベージュ)		○	

http://www.sharp.co.jp/sc/seihin/denka.html

図 4.2: 製品一覧表の例

ならない。このような場合は、製品ページ自体が持つ以下の特徴を利用することにより、ページに得点付けを行い、得点の高いページを製品一覧ページとして収集する。

- タイトル、逆リンク[1]のアンカ文字列、ページ内の強調された文字列、に「仕様」等の特徴的な語句、カテゴリ名が用いられている。
- 製品仕様表中にカテゴリ名が存在する。

製品情報抽出のアルゴリズムを以下に示す。

1. 製品一覧表の探索
   (a) 3章で述べたカテゴリ一覧表収集と同様の方法で製品一覧表収集を行う。但し、収集対象とする範囲は、カテゴリ URL から深さ１（1 リンク先）までとする。また、一覧表の判定では、カテゴリ名の代わりに型番を使用する。
   (b) 製品一覧表が収集できた場合は、型番とリンクを抽出して終了する。
2. 得点付けによる製品ページ収集
   (a) カテゴリ URL から 2 リンク先までのページを全て収集する。
   (b) 各ページに対して、以下の方法で得点をつける。
      - タイトル
        ページのタイトルはそのページの内容の要約となっている場合が多い。そこで、タイトルに以下に示すような特徴的な表現が含まれている場合、そのページはカテゴリ名に関して書かれている製品ページである可能性が高いとみなし、得点を 15 点加算する。

        仕様、スペック、SPEC、性能、定格、カタログ、一覧、ラインアップ、カテゴリ名、「型番のパターン」

      - 逆リンクのアンカ文字列
        リンク中のアンカ文字列は、リンク先ページの内容を表していることが多い。そこで、逆リンクのアンカ文字列に上記の表現が含まれている場合、そのページが製品ページである可能性が高いとみなし、得点を 13 点加算する。

[1]他のページから対象ページへのリンク

- 強調された文字列
  4.3.3節で述べる得点付けの方法と同様の方法を用い、ページ内で強調されている文字列を抽出し、それぞれに上記した特徴的な表現が含まれる場合、抽出した各文字列の得点を加算する。
- スコアの高いページからのリンク
  上記した 3 つの方法で得点付けを行った後、収集したページの得点の平均点を求める。そして、「信頼度の高いページからリンクされているページも信頼度が高い」という考えに基づき、、平均点よりも高い得点を持つページからリンクされているページに 13 点加点する。

(c) もう一度、平均点を計算し直し、平均点よりも高い得点を持つページを製品ページとする。

4.3 製品情報抽出

製品情報抽出の第 2 ステップは、製品ページからの製品情報抽出である。製品情報抽出では、収集した製品ページから、仕様表解析、ページ見出し解析、強調文字抽出の 3 つの方法により、製品情報を抽出する。

4.3.1 仕様表解析

多くの製品において、その製品の主要な情報は、仕様表という形で記述されるのが一般的である。仕様表の例を図 4.3に示す。

ウェブの製品ページにも、多くの場合、仕様表が掲載されている。これらの仕様表は、ほとんどの場合、テーブルタグを用いて記述される。このため、このような仕様表をテーブルタグを利用して解析することにより、仕様表からその製品の主要な情報を抽出することができる。仕様表解析は津田 [3] や山本 [4] によるテーブル解析の手法を参考に作成した。

図 4.3は最も基本的な仕様表の例である。この表は以下の条件を満たしている。

- 一列目にテーブルヘッダ[2]が存在する。
- 一列が一レコードになっている。
- セルの結合が存在しない。

[2]表見出し。データの属性を規定する。

このような条件を満たす表を「標準形」と定義する。仕様表が標準形であれば、そこから製品情報を抽出すること、すなわち、テーブルヘッダ（属性）とそれに対応するデータ（属性値）の組を抽出することは、非常に容易に実現できる。

しかしながら、ウェブ上に存在する仕様表は、必ずしも標準形となっているわけではない。例えば、次のような形式の表が存在する。

- セルが結合している。
  セルの結合とは、2 つ以上のセルにまたがって、データが掲載されている状態を指す。例えば、図 4.4では 1 行目の「献立アドバイス液晶レンジ」が、2 つのセル（2 列目と 3 列目）にまたがって掲載されている。このような状態をセルが結合している、と呼ぶことにする。

- テーブルヘッダが複数列存在する。
  表の中には図 4.4のように複数の列を使ってテーブルヘッダを掲載している表が存在する。この図では 1 列目と 2 列目がテーブルヘッダの列である。

- 見出し行が存在する
  見出し行とは１つの行の全てのセルが結合している行のことを指す。図 4.4の 1 行目は「製品仕様」という 1 セルのみなので、見出し行である。

- テーブルの方向が逆テーブルの方向とは、1 レコードの向きを指す。1 レコードが 1 列になっている表を縦向き、1 行になっている表を横向きと呼ぶことにする。標準形の表は縦向きであるが、横向きの表も存在する。

上記のような表から、製品に関する情報を属性と属性値の組として抽出することは、それほど単純ではない。そこで、まず、テーブル解析を行って非標準形の表を標準形に変換し、その後、情報を抽出するという方法を採用する。

仕様表解析のアルゴリズムを以下に示す。

1. 表を標準形に変換する
   表の標準化は、結合セルの解除、表の方向の標準化、テーブルヘッダ列の標準化、により行う。

   (a) 結合セルの解除
   HTML において td タグや th タグの中でセルの結合に用いられる colspan 属性、rowspan 属性を手がかりにセルの結合を解除し、同じ値を持つ複数のセルに分割する。図 4.5におけるセルの結合を解除した表が、図 4.6である。

型　式	C21-VT7B	C14-VT7B
消費電力	90W(待機時5.5W)	72W(待機時5.5W)
年間消費電力	153kW・h/年	120kW・h/年
外形寸法 (幅×高さ×奥行)	52.0×50.5×46.5cm	41.5×39.5×37.5cm
質　量	24.3kg	12.6kg

http://www.hitachi.co.jp/Prod/cpim/hkji0403.htm

図 4.3: 基本的な仕様表（標準形）

製品仕様			
品 名		献立アドバイス液晶レンジ	
形 名		RE－M210	RE－M110
定格電圧(100V)		50／60Hz共用	50／60Hz共用
レンジ加熱	定格消費電力	1，420W	1，280W
	高周波電力	900W、500W、200W相当	650W、500W、200W相当
オーブン加熱	定格消費電力	1，380W	1，400W
グリル加熱	定格消費電力	1，330W	1，400W
トースター加熱	定格消費電力	1，380W	1，400W
外形寸法		520×485×345	490×395×320
加熱室有効寸法		325×350×215	305×325×175
庫内容量		30L	22L
ターンテーブル直径 (セラミック皿)		33cm	31cm
質 量		約19kg	約17kg
オーブン温度調節範囲		40℃(発酵)・110～250℃※	40℃(発酵)・110～250℃
カラー		シルバー(S)	シャインベージュ(C) ホワイト(W) ダークグレー(H)

http://www.sharp.co.jp/sc/gaiyou/news/990615.html

図 4.4: 複雑な仕様表

<table>
<tr><td colspan="2">品名</td><td colspan="3">全自動洗濯機</td></tr>
<tr><td colspan="2">愛称</td><td colspan="3">パワー速洗力</td></tr>
<tr><td colspan="2">形名</td><td>ES-SE91</td><td>ES-SE81</td><td>ES-SE71</td></tr>
<tr><td colspan="2">標準価格 (円}</td><td>130,000</td><td>120,000</td><td>112,000</td></tr>
<tr><td rowspan="2">消費電力
（Wh）</td><td>稼働時</td><td>117</td><td>109</td><td>105</td></tr>
<tr><td>待機時/日</td><td colspan="3">0</td></tr>
<tr><td colspan="2">月産 (台）</td><td colspan="3">20,000</td></tr>
</table>

図 4.5: セルが結合している表

品名	品名	全自動洗濯機	全自動洗濯機	全自動洗濯機
愛称	愛称	パワー速洗力	パワー速洗力	パワー速洗力
形名	形名	ES-SE91	ES-SE81	ES-SE71
標準価格 (円)	標準価格 (円)	130,000	120,000	112,000
消費電力 (Wh)	稼働時	117	109	105
消費電力 (Wh)	待機時/日	0	0	0
月産 (台）	月産 (台）	20,000	20,000	20,000

図 4.6: セルの結合を解除した表

品名	全自動洗濯機	全自動洗濯機	全自動洗濯機
愛称	パワー速洗力	パワー速洗力	パワー速洗力
形名	ES-SE91	ES-SE81	ES-SE71
標準価格 (円)	130,000	120,000	112,000
消費電力 (Wh)	117(稼働時)	109(稼働時)	105(稼働時)
消費電力 (Wh)	0(待機時/日)	0(待機時/日)	0(待機時/日)
月産 (台）	20,000	20,000	20,000

図 4.7: テーブルヘッダ列の標準化後

(b) 表の方向判定

領域知識内のテーブルヘッダリストを用いて、表の方向判定を行う。表の 1 行目と 2 行目、1 列目と 2 列目に含まれているテーブルヘッダの数を数え、表の方向を調べる。標準形はテーブルヘッダが縦に並んだ表である。これを縦向きと呼ぶ。もし、表が横向きだった場合、行と列を入れ替え、表の向きを変える。

(c) テーブルヘッダ列の標準化

領域知識内のテーブルヘッダリストを用いて、テーブルヘッダ列の範囲を調べ、複数列だった場合は一列にまとめる。例えば、図 4.6では 1 列目と 2 列目がテーブルヘッダ列になっている。このうち、1 行の内容が同じもの（図 4.6中では品名、愛称、形名、標準価格、月産）は一つにまとめる。1 行の内容が異なるもの（消費電力）はテーブルヘッダリストと一致するものを残し、一致しないものは括弧をつけてデータフィールドに入れる。図 4.6の表にテーブルヘッダ列の標準化を行った結果を図 4.7に示す。

2. 標準化した表からデータ抽出を行う

表中の製造番号別に、テーブルヘッダ名とそれに対応するデータを組にしたものを製品情報として抽出する。これは以下の手順で行う。

(a) テーブルヘッダ列から型番等のテーブルヘッダの位置を取得する。

(b) 各レコード毎にテーブルヘッダとデータの組を抽出する。

(c) 型番を識別子としてデータを整理する。

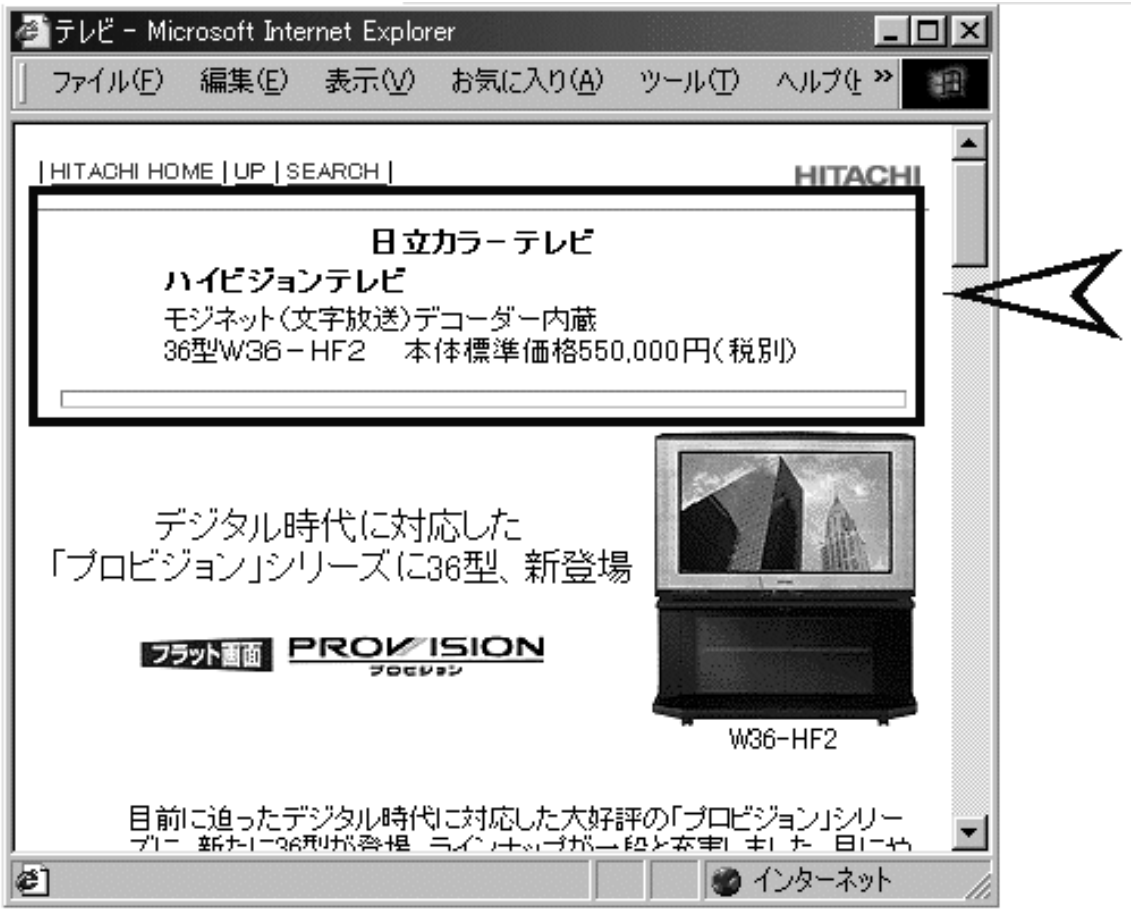
4.3.2 ページ見出し解析

ページ見出し解析では、領域知識内の共通フィールドと特徴フィールドを用い、ページ見出しから製品の主要属性を抽出する。

典型的なページ見出しの例を図 4.8の黒枠で囲った部分に示す。

この図に示すように、ページ見出し中の製品情報は仕様表の場合と異なり、明示的に製品属性が付与されていないことが多い。そこで、ページ見出しからの製品情報抽出では、領域知識内のテーブルヘッダのパターン（正規表現）を用い、パターンマッチングによって主要属性の抽出を行う。

ここでは、図 4.8のページ見出しを例に説明する。まず始めに、共通フィールドである型番、価格、品名、の抽出を試みる。領域知識内に記述している型番、価格、品名のパターンを以下に示す。



http://www.hitachi.co.jp/Prod/cpim/hkj01027.htm

図 4.8: ページ見出しの例

```
型番 (?:^|[^\dA-Z])([\dA-Z]{0,4}-[\dA-Z]{1,6})(?:$|[^\dA-Z])
価格 (?:(?:価格.{0,5}?|\s)?([1-9](?:[\d]{1,3}[,\.]?){1,3}.{1,3}円)|
          (?:(オープン|open)(?:価格)?)|
          (?:\\)\s?([1-9](?:[\d]{1,3}[,\.]?){1,3})
品名 無し
```

このうち品名はパターンが記述されていないので抽出を行わない。

次に、このページは製品カテゴリ「テレビ」の製品ページなので、特徴フィールドとしてはサイズの抽出を試みる。サイズの抽出に用いるパターンを以下に示す。

```
(?:\s｜大きさ｜サイズ)(\d{0,5})(型｜インチ)
```

これらを用い、パターンマッチングを行った結果として、抽出できる製品情報を以下に示す。

型番： W36-HF2、価格： 550,0000、品名：なし、サイズ：36 型

4.3.3 強調文字列抽出

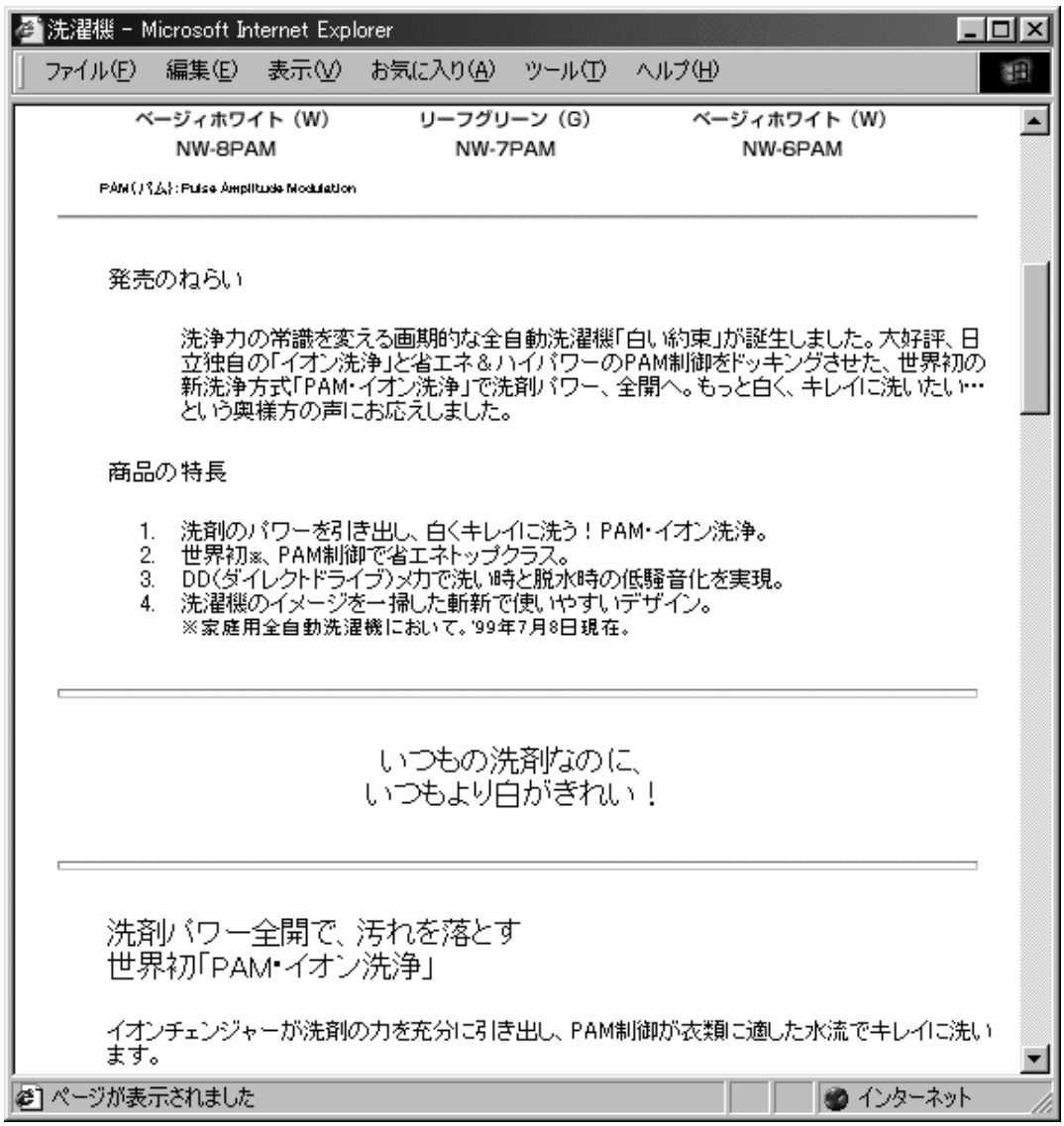
強調文字列抽出では、文字列を強調するタグを手がかりに、製品ページ内から製品のセールスポイントを抽出する。

典型的な製品ページの例を図 4.9に示す。

文字列を強調するタグ（強調タグ）は以下に示す 3 種類に分類することができる。

- 文字の大きさを変化させる 、<h1> から <h6>、<big>
- 文字の太さやフォントを変化させる <b>、、
- 文字列の先頭に記号などを付与し字下げを行う 、<ol>、<dl>
  これらは、文字列自体を直接変化させるわけではないが、他の文字列と差別化を図るという意味では文字列を強調しているといえる。

強調文字列抽出ではこれらのタグを利用して、強調されている文字列に強調の度合いによって得点付けを行い、得点の高い文字列を強調文字列として抽出する。このためには、強調タグとそのタグの範囲内にある文字列を一緒に抽出しなければならないが、強調タグ



http://www.hitachi.co.jp/Prod/cpim/hkj0701d.htm

図 4.9: 製品ページ例

は入れ子になっていることが多く、単純なパターンマッチングで強調タグを抽出することは難しい。

そこで、強調タグの有効範囲を明確にするため、まず入れ子になっている強調タグの分割を行う。分割を行えば、簡単なパターンマッチングにより強調タグとそのタグの範囲内にある文字列を一緒に抽出することができる。

強調文字列抽出の具体的な手順を以下に示す。

1. 強調タグの有効範囲を明確化する。
   入れ子になっている強調タグを分解する。例として、以下のような HTML コードを考える。

    だから、 いつもよりも白がきれい！ に洗えます。
   これは当社独自のイオン… … …

   この例では、文字列の中で、ある一文だけを強調するために の範囲内で更に別の を使用している。このような場合、以下に示すように外側の を分割し、有効範囲を明確にする。

    だから、
    いつもよりも白がきれい！
    に洗えます。 これは当社独自のイオン… … …

2. 強調タグを抽出する。
   有効範囲が明確になった強調タグとその範囲内にある文字列をパターンマッチングにより抽出する。

3. 得点付けを行う
   得点付けは上記した強調タグの種類それぞれに対して、以下のように行う。

   - 文字の大きさを変化させるタグ は以下の式により得点を決定する。

     文字サイズ× 2

     <font size=x> については x をそのまま文字サイズとする。ただし、<font size=+x>、<font size=-x> については相対的に文字の大きさを規定するタグであるため、標準の文字サイズ 3 からの差分を求め文字サイズとする。例えば、

<font size=+1> の場合は文字サイズ 4、<font size=-2> の場合は文字サイズ 1、とする。

<h1> から <h6> は <font> の場合と逆で、<hx> の x が大きくなるほど表示される文字サイズは小さくなる。この場合、文字サイズは 7-x により求める。また、<big> に関しては <font size=+1> と同等とする。

- 文字の太さやフォントを変化させるタグは得点に 1 点加える。また、<h1> から <h6> は大きさを変化させるだけでなく文字を太字に変える効果を持っている。そこで、<h1> から <h6> に関しても得点に 1 点加える。
- 文字列の先頭に記号などを付与し字下げを行うタグは得点に 2 点加える。

4. 得点付けを行った文字列を得点順にソートする。同じ得点になった場合は、掲載順にソートする。ソートした結果から上位 n 件（現在は 3）を強調文字列として取得する。ただし、抽出した文字列が以下の正規表現にマッチする場合、抽出しても意味のない文字列と判断し抽出しない。

仕様 | 特徴 | 特長 | 別売り | オプション

このような処理の結果、図 4.9からは、

- いつもの洗剤なのに、いつもより白がきれい！
- 洗剤パワー全開で、汚れを落とす世界初 PAM・イオン洗浄
- 洗剤のパワーを引き出し、白くキレイにあらう！PAM・イオン洗浄

という文字列が強調文字として抽出される。

第 5 章

評価実験

本章では、本研究で作成したカテゴリ別製品一覧表生成システムの有効性を検証するため評価実験を行った。

5.1 カテゴリ一覧表収集実験

カテゴリ URL 収集では、カテゴリ一覧表の収集が中心課題になる。カテゴリ一覧表を収集することができれば、カテゴリ URL の抽出は比較的簡単に行うことができる。このため、ここではカテゴリ一覧表を収集できるかどうかについて実験を行う。

家電製品を製造する 5 メーカを対象として、カテゴリ一覧表収集実験を行った。

評価方法は、人手によってカテゴリ一覧表を収集し、プログラムによる出力と比較した。

比較結果を以下のように分類する。

○ 人手と一致

人手で収集したカテゴリ一覧表と一致したもの。

人手と不一致

人手で収集したカテゴリ一覧表と一致しない場合、以下の 2 つに分類する。

△ 人手でも判断不可能なもの

メーカサイトの中には、図 5.1の例のようにカテゴリ一覧表と同じ特徴をもつ表が存在する。

× カテゴリ一覧表ではない

メーカ名	人手による抽出	プログラムによる抽出			
		○	△	×	計
日立製作所	6	6	0	0	6
シャープ	1	1	4	0	5
SONY	4	4	1	2	7
松下電器	8	8	6	2	16
東芝	2	2	1	2	5
ビクター	1	1	4	1	6

表 5.1: カテゴリ一覧表収集実験結果

以上のような分類を行った結果を表 5.1に示す。

人手で収集したカテゴリ一覧表を全て収集することができた。また、本モジュールでは、表層的な特徴のみを頼りに、カテゴリ一覧表かどうかの判断を行うため、△に分類したような表も収集してしまう。しかし、このような表はリンク先から製品仕様が収集できないので、製品情報収集の段階で無視され、精度には直接的に影響しない。

5.2 製品情報収集実験

製品情報収集は、製品ページ収集と製品情報抽出から構成されている。このため、製品ページ収集実験と、製品情報抽出実験を行った。

5.2.1 製品ページ収集実験

製品ページ収集実験では、5.1節で抽出したカテゴリ URL を用いて実験を行った。

収集対象は、家電製品を製造している 5 メーカのサイトの中から 2 カテゴリ（テレビ、掃除機）の製品ページとした。ただし、掃除機は 2 つのメーカで製造されていなかった。

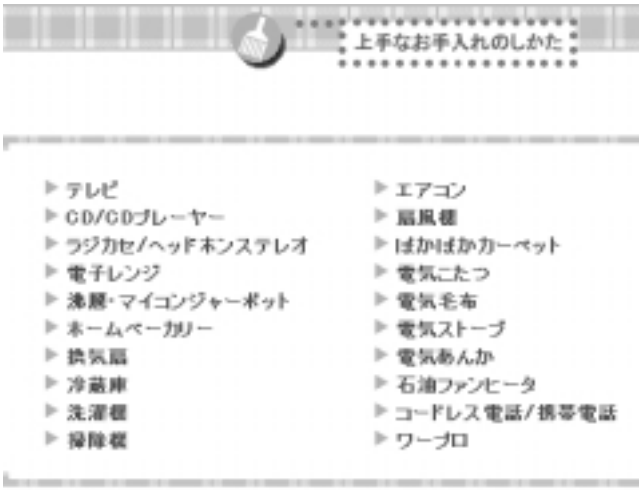
表 5.2に結果を示す。この表の評価基準は以下の通りである。

◎ 人手と一致

人手で収集した製品ページと一致した数

人手と不一致

人手で収集した製品ページと一致しない場合、以下の 3 つに分類する。



http://www.panasonic.co.jp/corp/css/oteire.html

図 5.1: カテゴリ一覧ではない表

○ 製品ページ
　人手では収集できなかった製品ページを収集した数

△ 製品ページではないページ
　製品ページではないページを収集した数

× 他の製品カテゴリのページ
　他の製品カテゴリのページを収集した数

表 5.2に示すように、各メーカによって多少の差はあるが、人手で収集した製品ページのほとんどを収集することができた。○の項目は人手では収集できなかったページ数だが、これは、1 つの製品に対する製品ページが複数存在する場合に、見逃してしまったものと考えられる。△は製品ページではないページを収集してしまった場合である。これには以下の 2 つの原因が考えられる。

- 製品一覧表のリンク先が製品ページではなかった
  製品一覧表のリンク先からもう 1 リンク辿らなければ製品ページにたどり着かない場合がある。

- 得点付けにより収集された
  得点付けによる製品ページ収集では、ページタイトルやアンカ文字列を手がかりと

表 5.2: 製品ページ収集実験結果

メーカ名	人手による抽出	プログラムによる抽出				
		◎	○	△	×	計
日立製作所	15	15	0	1	0	16
シャープ	29	24	0	0	4	28
SONY	38	35	3	8	1+4*	51
松下電器	61	51	28	27	0+7*	96
ビクター	11	11	19	26	0	56

*:名前からは見分けがつきにくい、他カテゴリの製品。
（スカイパーフェク TV 受信セット、エアクリーナー等）

するため、実際には製品ページではないページにも高い得点を与えてしまうことがあり、そのページを製品ページだと判定してしまう。

△に分類されたページからは製品情報を抽出することができないため、実際には精度とは関係がない。

×は他の製品カテゴリの製品ページを収集してしまった場合である。このようなページからは間違った製品情報が抽出されてしまうため、精度を低下させる直接の要因となる。原因としては、得点付けによる製品ページ収集で、収集している製品カテゴリと共通点が多い製品カテゴリのページにも高い得点が与えられてしまうことが考えられる。例えば、ビデオデッキはテレビと関係の深い製品カテゴリなので、ビデオデッキのページでも「テレビ」という表現が多用されることがある。このような場合、そのページに「テレビ」として高い得点を与えてしまい、製品ページだと判定してしまう。

5.2.2 製品情報抽出実験

製品情報抽出実験では、製品ページ収集実験の結果を用いて製品情報抽出を行った。

抽出対象は 4 メーカの 3 カテゴリ（テレビ、洗濯機、掃除機）の製品で、人手によって抽出した製品情報とプログラムによる出力を比較した。表 5.3に結果を示す。

この表の評価基準は以下の通りである。

- 型番
  - ○ 人手による抽出と一致した

表 5.3: 製品情報抽出実験結果

メーカ名	人手による抽出	プログラムによる抽出							
		抽出数	型番				価格		
			○	△	×	抽出率	○	×	抽出率
日立製作所	32	40	31	3	1	97%	30	1	94%
シャープ	50	87	50	32	5	100%	49	0	98%
松下電器	44	41	33	4	4	75%	16	17	37%
SONY	35	84	31	42	7	89%	29	35	82%

  - △ オプション、または過去の製品を抽出した
  - × 他のカテゴリ、関係ない文字列、または不完全な型番を抽出した

- 価格
  - ○ 正しい価格
  - × 間違った価格、または得られなかった

表 5.3に示すように、製品（型番）は全体で 90%、主要属性は全体で 77%の抽出率を得ることができた。これはメーカサイトに、グラフィックデータを多用している、バラエティーに富んだレイアウトを含んでいる、など機械化を妨げる要素が多いことを考慮すると良い結果であると言える。

製品情報抽出が失敗した原因のほとんどは、ページ見出しの抽出範囲が間違っていることが原因だった。本システムでは、ページ見出しは hr タグや br タグを手がかりに抽出を行う。しかし実際には table タグを使って書かれている変則的なものも存在し、それらを正しく抽出することができなかった。また、このような特徴はカテゴリ内で統一される傾向があるため、抽出できるレイアウトの場合は、そのカテゴリ内の製品のほとんどを抽出することができる反面、適用できないレイアウトが使用されている場合、ほとんどの製品を抽出できない、という極端な結果が見られた。

また、製品のオプション（テレビではリモコン、テレビ台など）を抽出してしまうことも問題点として挙げられる。これは、「フォントの小さな型番は抽出しない」というヒューリスティックで対処しているものもあるが、排除できないものも多く、根本的な解決策が必要である。

5.3 検討

本節では、前節までに行った実験結果をもとに、作成したシステムの有効性について検討を行う。

本研究では製品選択における意思決定を支援するために、製品情報の再組織化を行い、カテゴリ別製品一覧表を生成するシステムを作成した。カテゴリ別製品一覧表とは、特定カテゴリの製品に関する情報を提示した一覧表のことで、2.1節において以下の条件を満たしている必要がある、と定義した。

1. そのカテゴリの製品を網羅する（できるだけ多くの製品を含む)。

2. それぞれの製品に対して、その製品の主要情報を含む。

製品情報収集実験では、システムが生成するカテゴリ一覧表が、これらの条件を満たすことができるかを検証するため、人手によって製品情報を収集し、システムの収集結果との比較を行った。

まず、製品の網羅性としては、90%という結果を得ることができた。これは、製品ページ収集が有効に機能していることを示している。

また、主要属性の抽出としては、77%という結果を得ることができた。これはメーカサイトでは、グラフィックデータが多用される、バラエティーに富んだレイアウトが使用される、など機械化の妨げとなるような要素が多いことを考慮すると、本研究で提案した製品情報抽出が有効に機能していることを示している。

これらの実験結果から、上記の条件を満たすカテゴリ一覧表を生成することができ、製品選択における意思決定の際に有効な情報を提供するシステムを作成することができたといえる。

以下では、本研究で用いた各手法に関して、検討を行う。

- 製品ページ収集
  本研究で提案した製品ページ収集は、製品一覧表を利用する方法と、ページに得点付けを行う方法、を用いて行っている。製品一覧表は、製品ページへのリンク集であるが、情報収集を行う場合にリンク集を利用することが有効であることは Clever サーチ [5] や佐藤 [6] らのリンク集の自動生成、山本らの人物情報の自動収集 [4] によっても示されている。また得点付けによる製品ページ収集では、逆リンク情報、ページタイトル、ページ内の強調された文字を頼りに、製品ページかどうかの判定を行っているが、ページの内容を知るのに、このような情報が重要であることは大

槻 [7] によっても示されている。本研究では、これら 2 つの手法を組み合わせることによって、対象となるほとんどの製品ページを収集することに成功した。

- 製品情報収集
  本研究で提案した製品情報抽出は、仕様表解析、ページ見出し解析、強調文字抽出、を用いて行っている。製品仕様表の解析としては嶋田ら [8] の研究があるが、本研究ではこれに加え、ページ見出しからも製品情報の抽出を行い、これが有効に働くことを示した。

第 6 章

結論

本研究では製品選択における意思決定を支援するために、メーカサイト毎に存在する製品情報を製品カテゴリ毎に組織化し直し、カテゴリ別製品一覧表を生成してユーザに提供するシステムについて述べた。カテゴリ別製品一覧表とは、特定カテゴリの製品に関する情報を提示した一覧表のことで、以下の条件を満たしている必要がある。

1. そのカテゴリの製品を網羅する（できるだけ多くの製品を含む)。
2. それぞれの製品に対して、その製品の主要情報を含む

作成したシステムは、製品情報収集モジュール、実売価格調査モジュール、ユーザインタフェースモジュール、の３つのモジュールと、製品情報データベース、領域知識から構成される。製品情報抽出モジュールはメーカサイトからカテゴリ名を付与した形で製品情報を抽出し、製品情報データベースを作成する。収集した製品情報を用い、実売価格調査モジュールがオンラインショップから実売価格を調査し、製品情報データベースへ保存する。ユーザインタフェースモジュールでは、ユーザの要求を受付け、製品データベースから製品一覧表を生成する。

本システムの中核となる製品情報モジュールは、カテゴリ URL 収集モジュール、製品情報抽出モジュール、の 2 つのサブモジュールから構成されている。カテゴリ URL 収集モジュールでは、製品ページへのリンクがまとめられている表であるカテゴリ一覧表を収集し、その中からカテゴリ名とカテゴリ URL の抽出を行う。

製品情報抽出モジュールでは、カテゴリ URL を手がかりに製品ページを収集し、収集した製品ページから製品情報を抽出する。具体的には、カテゴリ URL を出発点として、製品一覧表を用いる方法、製品ページの特徴を利用し得点付けを行う方法、の 2 つの方法

により製品ページを収集する。収集した製品ページからは、仕様表解析、ページ見出し解析、強調文字抽出、により製品情報の抽出を行う。

最終的にシステム全体の評価実験として、家電製品を製造している 5 つのメーカにおいて、製品情報収集実験を行った。まず、カテゴリ URL 収集モジュールでは、人手で収集した 22 のカテゴリ一覧表全てを収集することができた。

製品情報収集モジュールでは、人手で収集した 161 の製品のうち 145 を収集することができ、抽出率は 90%だった。また、主要属性は 161 のうち 124 に対して、正確な値を抽出することができ、抽出率は 77%だった。これはメーカサイトに、グラフィックデータを多用している、バラエティーに富んだレイアウトを含んでいる、など機械化を妨げる要素が多いことを考慮すると良い結果であると言える。

これらの実験結果から、上記の条件を満たすカテゴリ一覧表を生成することができ、製品選択における意思決定の際に有効な情報を提供するシステムを作成することができたといえる。

今後の課題としては、カテゴリ一覧表の使いやすさに関して調査を行い、製品選択における意思決定に対しての有効性を検証する、ことが考えられる。

謝辞

本研究を進めるにあたり、多くの御教示、熱心な御指導を賜りました佐藤理史助教授に深く感謝致すとともに、心から御礼申し上げます。そして、日頃から技術的にも精神的にも支援して下さいました知識工学講座の皆様に感謝したいと思います。

最後に、大学院での 2 年間、様々な面で支えてくれた家族、友人に感謝します。ありがとうございました。

参考文献

[1] Robert B. Doorenbos, Oren Etzioni, and Daniel S. Weld, "A Scalable Comparision-Shopping Agent for the World Wide Web", In *Proceedings of the First International Conference on Autonomous Agents*, 1997.

[2] 冨田 一郎, 手塚 祐一, 山本 修一郎, 長岡 満夫, "HTML 文章からの商品情報抽出方式の提案", 情報処理学会全国大会第 56 回全国大会予行集 (3), pp.79-80, 1998.

[3] 津田 朋樹," ワールドワイドウェブからの住所録の自動生成", 北陸先端科学技術大学院大学修士論文, 2000

[4] 山本 あゆみ, 佐藤 理史, "ワールドワイドウェブからの人物情報の自動収集", 第 199 回情報処理学会「知能と複雑系」研究会（ICS-119）, pp173-180, 2000.

[5] Members of the Clecer Project, "Hypersearching the Web", Scientific American, Vol.280, No.6, pp54-60, 1999.

[6] Satoshi Sato and Madoka Sato, "Toward Automatic Generation of Web Directories", Proc of International Symposium on Digital Libraries 1999(ISDL'99), pp127-134, 1999.

[7] 大槻 洋輔, 佐藤 理史, "ワールドワイドウェブを知識源とした地域情報の自動編集", 第 119 回情報処理学会「知能と複雑系」研究会（ICS-119）, pp165-172, 2000.

[8] 嶋田 和孝, 遠藤 勉, "製品性能表からの特徴データの抽出", 第 133 回情報処理学会「自然言語処理」研究会（NL-133）, pp107-113, 1999.

Appendix A

領域知識

以下に本研究で設計した領域知識を記す。

```
<domain_description>
<maker_list>
<maker> <name>HITACHI</name>
<url>www.hitachi.co.jp/index-j.html</url>
<alias><li>日立 <li>ヒタチ </alias></maker>
<maker> <name>SHARP</name>
<url>www.sharp.co.jp</url>
<alias><li>シャープ</alias></maker>
<maker> <name>SONY</name>
<url>www.sony.co.jp</url>
<alias><li>ソニー</alias></maker>
<maker> <name>PANASONIC</name>
<url>www.panasonic.co.jp</url>
<alias><li>パナソニック <li>ナショナル <li>NATIONAL <li>松下 </alias></maker>
<maker> <name>SANYO</name>
<url>www.sanyo.co.jp</url>
<alias><li>三洋 <li>サンヨー</alias></maker>
<maker> <name>TOSHIBA</name>
<url>www.toshiba.co.jp</url>
<alias><li>東芝 <li>トウシバ</alias></maker>
<maker> <name>VICTOR</name>
```

```
<url>www.victor.co.jp</url>
<alias><li>ビクター </alias></maker>
</maker_list>

<category_list>
<category> <name><li>テレビ <li>TV </name>
   <field></field>
</category>
<category> <name><li>ビデオカメラ <li>ハンディカム</name>
   <field></field>
</category>
<category> <name><li>ビデオデッキ <li>ビデオ <li>ビデオプレーヤー</name>
   <field></field>
</category>
<category> <name><li>DVD プレーヤ <li>DVD デッキ</name>
   <field></field>
</category>
<category> <name><li>オーディオ <li>ラジカセ <li>コンポ </name>
</category>
<category> <name><li>CD デッキ <li>MD デッキ </name>
</category>
<category> <name><li>カーナビゲーション <li>カーナビ </name>
</category>
<category> <name><li>デジタルカメラ <li>デジカメ </name>
</category>

<category> <name><li>冷蔵庫 <li>冷凍庫 </name>
   <field><li>容量</field>
</category>
<category> <name><li>洗濯機 </name>
</category>
<category> <name><li>レンジ </name>
   <field><li>容量 <li>電力</field>
```

```
</category>
<category> <name><li>炊飯器 <li>電気釜 </name>
</category>
<category> <name><li>電気ポット <li>ポット</name>
   <field><li>容量 </field>
</category>
<category> <name><li>コーヒーメーカ </name>
</category>
<category> <name><li>ホットプレート </name>
</category>
<category> <name><li>食器洗い </name>
</category>
<category> <name><li>乾燥機 </name>
</category>
<category> <name><li>浄水器 </name>
</category>
<category> <name><li>換気扇 </name>
</category>
<category> <name><li>ガスコンロ <li>ガス調理 <li>ガステーブル </name>
</category>

<category> <name><li>エアコン</name>
</category>
<cagegory> <nemes><li>扇風機 </name>
</category>
<category> <name><li>ファンヒータ <li>電気暖房 </name>
</category>
<category> <name><li>掃除機 <li>クリーナ </name>
</category>
<category> <name><li>留守番電話 <li>コードレスホン </name>
</category>
<category> <name><li>携帯電話 </name>
   <field><li>質量 </field>
```

```
</category>
<category> <name><li>ファクシミリ <li>FAX </name>
</category>
<category> <name><li>空気清浄器 <li>空気清浄機</name>
</category>
<category> <name><li>除湿器 <li>除湿機 </name>
</category>
<category> <name><li>加湿器 <li>加湿機 </name>
</category>
<category> <name><li>ホットカーペット <li>電気カーペット</name>
</category>
<category> <name><li>電気シェーバ <li>電気カミソリ <li>シェーバ <li>電気剃刀 </name>
</category>
<category> <name><li>ドライヤ </name>
</category>
<category> <name><li>アイロン </name>
</category>
</category_list>

<common_field>
<field key=true> <header_name>型番 </header_name></field>
<field>  <header_name>価格 </header_name></field>
<field>  <header_name>品名 </header_name></field>
</common_field>

<table_header_list>
<table_header> <name><li>型番 <li>品番 <li>形名 <li>型名 <li>型式 <li>機種名 </name>
       <pattern>(?:^|[^\dA-Z])([\dA-Z]{0,4}-[\dA-Z]{1,7})(?:$|[^\dA-Z])</pattern>
</table_header>
<table_header> <name><li>品名 <li>種類</name>
</table_header>
<table_header> <name><li>価格 <li>定価 </name>
```

```
        <pattern>(?:(?:価格.{0,5}?|\s)?([1-9](?:[\d]{1,3}[,\.]?){1,3}.{1,3}
円)|
   (?:(オープン|open)(?:価格)?)|
   (?:\\)\s?([1-9](?:[\d]{1,3}[,\.]?){1,3}))</pattern>
        <demonination><li>円 </denomination>
</table_header>
<table_header> <name><li>寸法 </name>

</table_header>
<table_header> <name><li>サイズ <li>大きさ</name>
        <denomination><li>型 <li>インチ</denomination>
</table_header>
<table_header> <name><li>電力 </name>
        <denomination><li>W <li>ワット</denomination>
</table_header>
<table_header> <name><li>質量 <li>重量 </name>
        <denomination><li>g <li>グラム </denomination>
</table_header>
<table_header> <name><li>色 <li>カラー </name>
</table_header>
<table_header> <name><li>愛称 </name>
</table_header>
<table_header> <name><li>月産 </name>
        <denomination><li>台</denomination>
</table_header>
<table_header> <name><li>容量 <li>容積 </name>
        <denomination><li>l <li>リットル </denomination>
</table_header>
<table_header> <name><li>付属品 <li>オプション</name>
</table_header>
<table_header> <name><li>仕事率 <li>吸い込み <li>吸引 </name>
        <denomination><li>W <li>ワット </denomination>
</table_header>
```

<table_header> <name>
備考 </name>
</table_header>
<table_header> <name>
メーカ </name>
</table_header>
<table_header> <name>
接続端子 入力端子 </name>
</table_header>
</table_header_list>