自然言語処理 133－15
（1999．9．10）

製品性能表からの特徴データの抽出

嶋田 和孝　　遠藤 勉

大分大学工学部知能情報システム工学科
〒870-1192 大分県大分市旦野原700番地
E-mail:{shimada, endo}@csis.oita-u.ac.jp

あらまし 製品の機能などを記述した製品性能表は，多くのデータを含んでいるが，どれがその製品の特徴であるかは言及されていない．本論文では，HTML で記述されている製品性能表を解析し，それぞれの製品の特徴となるデータの抽出方法について考察する．まず，製品性能表の解析処理について述べ，その結果生成された表構造を用いて製品の特徴を抽出する方法について述べる．特徴データは，表層的な数値や文字列の比較および他の特徴データからの推定処理により抽出される．最後に実験を通じて本手法の有効性を検証する．

キーワード 製品性能表，表構造，表層表現，特徴データ

Extracting Important Data from Specifications

Kazutaka SHIMADA　　**Tsutomu ENDO**

Department of Computer Science and Intelligent Systems, Oita University
700 Dannoharu Oita, 870-1192 Japan
E-mail:{shimada, endo}@csis.oita-u.ac.jp

Abstract Specifications contain many data, but it is not clear which is the most important data in them. In this paper, we propose a method for extracting important data from specifications.
The proposed method analyzes specifications and generates table structure. Important data are extracted from it by comparing surface expressions such as numerals and characters. Experimental results show the effectiveness of our method.

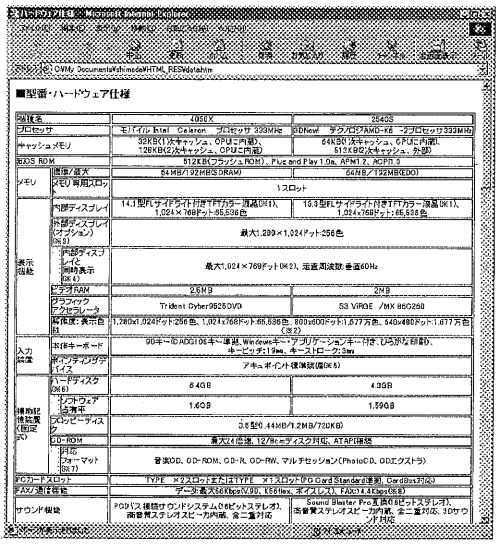
key words Specifications, Table structure, Surface expression, Important data

1．はじめに

近年の急速なネットワークの普及により，職場や家庭にいながら世界中から発信された情報にアクセスできる環境が整ってきた．これに伴い，今までは紙面で伝えられていた内容が電子化されてきている．その例として製品の機能等を記述した製品性能表（図１）というものが挙げられる．製品性能表には，その製品に関して多くのデータが記述されているが，逆に言えば記述されているだけで，どれがその製品の特徴なのかはわかりづらい．

従来，内容抽出に関する研究は，製品紹介の記事[1]や新聞記事[2][3]などの文書が対象とされることが多かった．また表のような文書からの内容抽出としては，会告記事に関する電子ニュースの自動要約[4]や罫線を含まない表形式文書を対象とした[5]などがあるが，その対象は箇条書きされた文書であり，一般に表と呼ばれるものとは多少異なり，WWW 上で公開されている HTML 文書については扱われていない．また，上記の内容抽出に関する研究は，処理される対象が１つの記事であり，複数の記事からの内容抽出は行われていない．同様に，複数の製品に対する相対的な製品特徴も抽出されていない．

本研究では，HTML で記述された製品性能表を解析し，その解析結果を用いて，複数の製品データからそれぞれの製品の特徴となるデータを数値と表層的な文字列の比較処理により推定する．


■型番・ハードウェア仕様

製種名	4050X	2540S
プロセッサ	モバイル Intel Celeron プロセッサ 333MHz	AMD-K6 -2プロセッサ 333MHz
キャッシュメモリ	32KB(1次キャッシュ、CPUに内蔵)、128KB(2次キャッシュ、CPUに内蔵)	64KB(1次キャッシュ、CPUに内蔵)、512KB(2次キャッシュ、外部)
BIOS ROM	512KB(フラッシュROM)、Plug and Play 1.0a、APM1.2、ACPI1.0	
メモリ 標準/最大	64MB/192MB(SDRAM)	64MB/192MB(EDO)
メモリ 専用スロット	1スロット	
表示機能 内部ディスプレイ	14.1型TFTサイドライト付きTFTカラー液晶(SXGA)、1,024×768ドット=65,536色	13.3型TFTサイドライト付きTFTカラー液晶(SXGA)、1,024×768ドット=65,536色
表示機能 外部ディスプレイ(オプション)(注3)	最大1,280×1,024ドット=256色	
表示機能 外部ディスプレイ同時表示(注4)	最大1,024×768ドット(XGA)、垂直周波数:最高60Hz	
表示機能 ビデオRAM	2.5MB	2MB
表示機能 グラフィックアクセラレータ	Trident Cyber9525DVD	S3 VIRGE /MX 86C260
表示機能 表示色数	1,280dot×1,024ドット=256色、1,024×768ドット=65,536色、800×600ドット=1,677万色、640×480ドット=1,677万色(注22)	
入力装置	本体キーボード	90キーのASCII配列キーボード搭載、Windowsキー・アプリケーションキー付き、ひろがる軽設計、キーピッチ:19mm、キーストローク:3mm
入力装置 ポインティングデバイス	アキュポイント(標準装備)(注5)	
補助記憶装置(固定式) ハードディスク(注6)	6.4GB	4.3GB
補助記憶装置(固定式) ソフトウェア当初モデル	1.6GB	1.59GB
補助記憶装置(固定式) フロッピーディスク	3.5型(0.44MB/1.2MB/720KB)	
補助記憶装置(固定式) CD-ROM 対応フォーマット(注7)	最大24倍速、12/8cmディスク対応、ATAPI接続	音楽CD、CD-ROM、CD-R、CD-RW、マルチセッション(PhotoCD、CDエクストラ)
PCカードスロット	TYPE Ⅱ×2スロットまたはTYPE Ⅲ×1スロット(PC Card Standard準拠、CardBus対応)	
FAX/通信機能	データ最大56Kbps(V.90、K56flex、ボイスレス3、FAX14.4Kbps(注8)	
サウンド機能	PCMなら機種サウンドシステム(16ビットステレオ)、両音声ステレオスピーカ内蔵、全二重対応	Sound Blaster Pro互換(16ビットステレオ)、両音声ステレオスピーカ内蔵、全二重対応、3Dサウンド対応

図１　製品性能表の例

2．表構造生成

対象としている表データは HTML で記述されているものとする．HTML では，表は行単位の集合として記述されている．複数の行や列にまたがるカラムはタグを用いて先頭の行や列だけにその情報が記述されている．図２(a)の表の HTML 表記が図２(b)である．

２次元で記述された表を１次元の記号列に変換する．これを表構造と呼び，以下のようなリスト形式で表す．

		PC1	PC2
CPU		P400MHz	P450MHz
メモリ	標準	64MB	128MB
	最大	256MB	
	VRAM	4MB	

(a)

```
<table border="1">

 <tr>
      <td colspan="2">  </td>
      <td>PC1</td>
      <td>PC2</td>
 </tr>
 <tr>
      <td colspan="2">CPU</td>
      <td>P400MHz</td>
      <td>P450MHz</td>
 </tr>
 <tr>
      <td rowspan="3">メ<br>モ<br>リ<br></td>
      <td>標準</td>
      <td>64MB</td>
      <td>128MB</td>
 </tr>
 <tr>
      <td>最大</td>
      <td colspan="2">256MB</td>
 </tr>
 <tr>
     <td>VRAM</td>
      <td colspan="2">4MB</td>
 </tr>
</table>
```

(b)

図２　HTML 表記の例

（列項目　行項目　データ）

列および行項目が複数存在する場合は，括弧を用いて以下のように記述する．

（（列項目 1....列項目 n）
（行項目 1....行項目 n）　データ）

表構造は以下の手順で生成される．

(1)項目とデータの対応付け

一般に製品性能表は階層的関係を持つ項目や複数の項目に対して共通のデータを持つ場合が多い．表構造は基本的にデータと項目を1対1の関係で記述するため，これらの複数の階層的な項目を分割し，それぞれのデータと対応づける必要がある．HTML における複数の列・行を記述するためのタグ（colspan および rowspan）を用いて，表の項目とデータの対応付けを行う．図3は図2(a)について項目・データ間の対応付け処理を行った結果である．図中の破線部分が補完された線である．この結果を用いて表構造に変換する．

(2)表構造への変換

表構造の”列項目”は基本的に製品名もしくは型名などが割り当てられる．列項目は，キーワード（製品名，型式，モデルなど）と位置情報を用いて推定する．図4は図3の表から生成された表構造である．

		PC1	PC2
CPU		P400MHz	P450MHz
メモリ	標準	64MB	128MB
メモリ	最大	256MB	256MB
メモリ	VRAM	4MB	4MB

図3　対応付けされた表

(PC1 CPU P400MHz)
(PC1 (メモリ 標準) 64MB)
(PC1 (メモリ 最大) 256MB)
(PC1 (メモリ VRAM) 4MB)
(PC2 CPU P450MHz)
(PC2 (メモリ 標準) 128MB)
(PC2 (メモリ 最大) 256MB)
(PC2 (メモリ VRAM) 4MB)

図4　表構造の例

(3)データ処理

表中では，1つの項目に対して複数のデータが存在する場合がある．表構造は列・行項目に対してデータは1つのみと定義してあるので，得られたデータに対して分割処理を行う必要がある．その他にも，括弧表現や表の特有表現（○，×，－，[]など）が多く存在する．これらに対しても以下のような処理を行い，表構造を生成する．

(3-1)データの分割

キーワード（／ や ，）を用いてデータを分割する．それに伴い，表構造も複数に分割される．（図5(a)）

(3-2)括弧表現の解析

括弧外要素と括弧内要素との関係を表層的なキーワードを用いて解析する．解析された結果は，表構造中のデータに付加される．（図5(b)）

(PC1　付属品　マウス，バックアップ CD....)
↓ **分割**
(PC1　付属品　マウス)
(PC1　付属品　バックアップ CD)

(a)

32KB(CPU に内蔵)
↓ **関係特定**
(32KB (state CPU 内蔵))

(b)

(PC1　スロット[空き]　1[0])
↓ **分割**
(PC1　スロット　1)
(PC1　(スロット 空き)　0)

(c)

図5　データに対する処理の例

(3-3)特殊表現への処理

表で用いられる特有表現に対して正規化，表構造の分割を行う．（図5(c)）

3．製品特徴抽出

製品性能表は，記述されている製品に対して多くのデータを含んでいるが，何がその製品の特徴であるかがわかりづらい．また，製品について詳しく知らないユーザにとっては，どれがその製品の特徴であるかを推測することが困難である．本節では，複数の製品についての製品性能表のデータから数値および表層的な文字列の比較により，各々の製品の特徴を抽出する手法について述べる．

3.1 最上位機種の特定

製品性能表では，一つの表に複数の製品に関する情報が記述されることが多い．そこで，まず，複数の製品の中からの最上位機種の推定法について述べる．これは，最上位機種の性能・装備が特徴データになる可能性があるからである．最上位機種はあらかじめ定めた基準項目に関するデータを比較することで推定する．基準項目としては，以下の3つが用いられる．

(1)CPUのクロック数
(2)ハードディスクの容量
(3)標準・最大メモリの容量

これら基準項目に対するデータに有意な差がなければ，以下の項目から最上位機種を推定する．

(4)表示機能に関する性能値
(5)機種・モデル名

機種・モデル名に関しては，同一シリーズでは，機種・モデル名中で使用される数値が大きいほど，もしくは末尾などに付属するアルファベットがZに近いほど上位機種の可能性が高いためである．

以上の基準項目に関するデータを比較し，最上位機種を特定する．特定処理における基準項目の優先順位もこれに準ずる．

3.2 数値によるデータの特徴推定

次に数値データを用いた特徴推定法について述べる．特徴データは，あらかじめ与えたキーワード（単位）に付属する数値を比較することで推定される．用いられるキーワードとしては，”円”，”MHz”，”MB”，”色”，”mm”，”W”などがある．これらキーワードを数値が大きいものほど高性能なものと小さいものほど高性能なものに分類し（表1），比較を行う．比較の方法としては単純に大小を比べる場合と複数の製品の同一項目のデータか

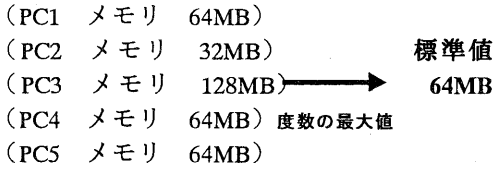


図6　標準値の算出例

表1　単位による数値の示す性能

数値が大きいほどよい単位	MHz，KB，MB，GB，色，ドット，型
数値が小さいほどよい単位	W，円，kg
項目によって変わる単位	時間，mm

表2　標準値の算出方法の分類

平均値	W，円，mm，kg，時間
度数の最大値	MHz，KB，MB，GB，色，ドット，型

ら標準値を求め，それを基準に比べる場合の２種類がある．標準値の算出方法は，キーワードおよび項目により異なり，”円”や”時間”など連続的に増減するものは，全データの平均値を求め，それを標準値とする．また，”MB”や”色”のように増え方に一定のパターンがあるものは，全データに対する度数の最も高いデータを標準値とする（表２）．得られた標準値を用いて比較を行い，その結果が標準値以上もしくは以下であれば，データが表中の最高値でなくとも，それを特徴として抽出する．

3.3 文字データの特徴推定

グラフィックボードやその他搭載されている機器に関して，それら機器名やそれに関する情報全てを辞書に記述しておくことは現実的に困難である．そこで，数値以外のデータに関しては，基本的には辞書を用いず，単純な文字列の比較により特徴推定を行う．各々の製品が異なるデータを持っている場合は，最上位機種のデータを特徴化する．また，基準項目に関するデータが同じ場合は，それぞれのデータを特徴化する．

しかし，単純な文字列の比較では十分に有用な特徴データを推定できるとは限らない．そこで，**3.2** で推定された特徴データから他のデータについて特徴を推定する．標準値より大きい値（もしくは小さい値）を持つ項目が階層構造を持つ場合，階層中にある他のデータも特徴化する．階層構造には，表があらかじめ持つ階層関係と知識として与える階層関係の２種類がある．

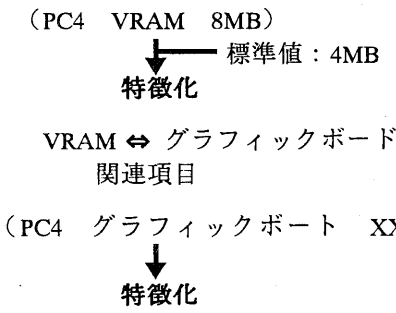


図７　項目の階層関係による特徴化

4．実験・考察

3．で示したアルゴリズムを基に実際の製品性能表を用いた評価実験を行った．以下，順にその結果に対して考察を行う．

まず，最上位機種の特定処理についてである．最上位機種の特定は前述したように，数値や知識辞書だけでは特徴化できない文字列データに関しての特徴データの抽出に必要である．実際の製品の中で，どれが最上位機種であるか，つまりどれが正解データであるかの判断は，著者の主観により決定した．4 シリーズ 12 機種に対して実験を行ったところ，全てのシリーズに対して最上位機種の特定に成功した．最上位機種の特定処理に関しては，処理対象となるほぼ全ての表中データが，数値であり，言語的な情報にあまり左右されないことが高い成功率の理由だと考えられる．実際には，特定処理では知識として幾つかの情報を用いているが（ROM の種類など），これらは，データ中に存在する他の文字列データと違い，網羅しやすいものであり，今後も知識の追加は必要となるであろうが，その数はそれほど急激には増加しないと考えられる．

次に各々の製品に対する特徴データの抽出処理について考察する．10 機種の製品性能表を種類毎に３つに分け実験を行った．その例を図８に示す．図８の製品性能表から特徴化されたデータの一覧を表３に示す．抽出されたデータが正しいかどうかの判断は，実際の雑誌のレビューを基準にして行った．雑誌では，PC2 に対する PC1 の製品特徴として以下の３点が挙げられていた．

1. 基本性能が高い
2. 操作性がよい
3. モバイル度が高い

PC1 に対する特徴化されたデータとして，「表

示ディスプレイ」と「VRAM」が抽出されている．これら表示機能は，最上位機種特定処理でも用いられたように，製品の基本性能を比べる際に重要な項目である．これにより，PC1 の製品特徴として抽出された「表示ディスプレイ」と「VRAM」は，正解であると言える．また，”モバイル度”に関しても，それに関連する「サイズ」，「重量」および「バッテリー駆動時間」が抽出されており，これらも正しく特徴化されたと言える．しかし，”操作性”については，PC2 の特徴データとして「キーピッチ」が抽出されており，雑誌のレビューとは一致しない．これは，双方のパソコンのポインティングデバイス，およびキーの配列，主要キーの大きさなどから，PC1 の方が”操作性がよい”と判断されたためである．これらの要素は表中には存在しないため，本手法では正しく特徴化できない．また，PC1 に対する PC2 の製品特徴としては，以下の 1 点が挙げられている．

1. 拡張性が高い

これに関しても，PC2 の特徴化されたデータとして，「メモリ（最大）」，「PC カードスロ

		PC1	PC2
CPU		300MHz	300MHz
2 次キャッシュ		128KB	128KB
メモリ	標準	64MB(SDRAM)	64MB(SDRAM)
	最大	128MB(SDRAM)	192MB(SDRAM)
表示ディスプレイ		1024 × 768 ドット	800 × 600 ドット
VRAM		2.5MB	2MB
PC カードスロット		Type Ⅱ × 1	Type Ⅰ/Type Ⅱ × 1
インターフェース		USB × 1，IrDA，iLINK 端子，ディスプレイアダプタ用端子	USB × 1，IrDA，シリアル，パラレル，ディスプレイ端子，マウス・キーボード共有ポート
サイズ		259mm(W)× 208mm(D)× 19.8mm(H)	260mm(W)× 202mm(D)× 26.6mm(H)
重量		約 1.2kg	約 1.32kg
バッテリー駆動時間		1.5 ～ 2.5 時間（標準），約 6.0 ～ 10.0 時間（最大）	約 1.9 時間（標準），約 4.0 時間（最大）
キーピッチ		17mm	18mm

図 8　実験に用いた製品性能表の例

表 3　特徴化されたデータの一覧

PC1 の特徴データ		PC2 の特徴データ	
表項目	データ	表項目	データ
表示ディスプレイ	1024 × 768 ドット	メモリ（最大）	192MB(SDRAM)
VRAM	2.5MB	PC カードスロット	Type Ⅰ/Type Ⅱ × 1
サイズ	259mm(W)× 208mm(D)× 19.8mm(H)	インターフェース	シリアル，パラレル
重量	1.2kg	キーピッチ	18mm
バッテリー駆動時間	1.5 ～ 2.5 時間（標準） 約 6.0 ～ 10.0 時間（最大）		

ット」および「インターフェース」が抽出されており，特徴化されたデータは正しかったと言えるだろう．

以上，PC1 の製品特徴として挙げられた”操作性”に関しては，誤った特徴抽出が行われているが，その他の抽出結果は妥当なものであり，本手法の有効性が確認された．

5．おわりに

本稿では，HTML で記述された製品性能表を解析し，表構造を生成し，複数の製品データからそれぞれの製品の特徴となるデータを数値や表層的な文字列の比較処理により推定する手法を提案した．また，評価実験により本手法の有効性を確認した．

現在，ユーザからの入力質問文を用いた動的な特徴データの抽出処理への拡張を考えている[8]．また，これらの抽出結果を用い，表を文章として要約，出力するシステムの構築を進めている．

参考文献

[1]松尾比呂志，木本晴夫，”抽出パターンの階層的照合に基づく日本語テキストからの内容抽出”，情処学論，vol.36，no.8，pp.1838-1844，1995.

[2]江里口善生，木谷　強，”富田一般化 LR パーザを用いた情報抽出”，情処学論，vol.38，no.1，pp.44-45，1997.

[3]若尾孝博，”英語テキストからの情報抽出 MUC 第 6 回大会参加報告”，情処学自然言語研報，NL114-12，1996.

[4]佐藤　円，佐藤理史，篠田陽一，”電子ニュースのダイジェスト自動生成”，情処学論，vol.36，no.10，pp.2371-2379，1995.

[5]河合敦夫，塚本雄之，山本勝紀，椎野努，”文書構造を利用した箇条書きや表形式文書からの内容抽出”，信学論（D-Ⅱ），vol. J81-D2，no. 7，pp1609-1619，1998

[6]嶋田和孝，織田雅也，賀川経夫，大城英裕，遠藤　勉，”ネットワーク環境における文書検索のための言語処理方式”，電気関係学会九州支部第 50 回連合大会論文集，1997.

[7]嶋田和孝，重松礼恵，織田雅也，遠藤　勉，”マルチメディア文脈処理のための表解析”，電気関係学会九州支部第 51 回連合大会論文集，1998.

[8]嶋田和孝，遠藤　勉，”表からの動的な特徴抽出・要約処理のための入力質問文の解析と分類”，電気関係学会九州支部第 52 回連合大会論文集，1999.