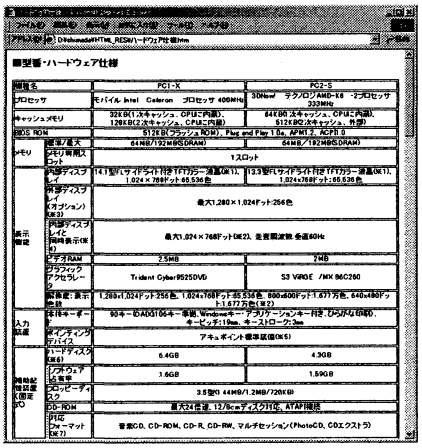
WWWからの製品性能表抽出

林 晃司 *1　　嶋田和孝 *2　　遠藤 勉 *2
*1 九州工業大学大学院情報工学研究科情報科学専攻
*2 九州工業大学情報工学部知能情報工学科

1 はじめに

インターネットならびにWWW(World Wide Web)の普及により，誰でも手軽に膨大な電子化文書へとアクセス出来る環境が整ってきた．しかしながら，WWWには多種多様の情報が共存する為，必ずしも欲しい情報のみを正しく検索・収集出来る訳では無い．また，仮に情報を収集出来たとしても，その量が膨大であれば閲覧は困難を極める．WWW上の情報検索においては単なる情報収集技術のみでは無く，収集した情報をいかに効率良く利用出来るかについても考慮されるべきである．

WWW上の情報の中でも，この要求に高い意義が持たれるものの一種に「製品性能表」がある．製品性能表とは製品の仕様書を電子化し，テーブル形式で記述したものを指す．図1に例を示す．



■型番・ハードウェア仕様

機種名		PC1-X	PC2-S
プロセッサ		モバイル Intel Celeron プロセッサ 400MHz	3DNow! テクノロジ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(プラグ&プレイ、PnP、APM1.2、ACPI 1.0	
メモリ	搭載/最大	64MB/192MBSDRAM	64MB/192MBSDRAM
	メモリ専用スロット	1スロット	
表示機能	内蔵ディスプレイ	14.1型TFTサイドライト付き(1677万カラー)液晶(XGA)、1,024×768ドット、65,536色	12.1型TFTサイドライト付き(1677万カラー)液晶(XGA)、1,024×768ドット、65,536色
	外部ディスプレイ(オプション)	最大1,280×1,024ドット・256色	
	内蔵ディスプレイと同時表示(オプション)	最大1,024×768ドット(XGA)、垂直周波数 垂直60Hz	
	ビデオRAM	2.5MB	2MB
	グラフィックアクセラレータ	Trident Cyber9525DVD	S3 VIRGE /MX 86C260
	液晶表示時	1,280×1,024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(XGA)	
入力装置	キーボード	86キー・ISO/JIS106キー準拠、Windowsキー・アプリケーションキー付き、ひらがな(106)、キーピッチ:19mm、キーストローク:3mm	
	ポインティングデバイス	アキュポイント標準装備(PS/2)	
補助記憶装置(固定I/O)	ハードディスク(SATA)	6.4GB	4.3GB
	インストールOS有り	1.6GB	1.59GB
	フロッピーディスク	3.5型(1.44MB/1.2MB/720KB)	
	CD-ROM	最大24倍速、12/8cmディスク対応、ATAPI接続	
	DVD(ファームウェアVer.7)	音楽CD、CD-ROM、CD-R、CD-RW、マルチセッション(PhotoCD、CDエクストラ)	

図 1: 製品性能表

ウェブドキュメントはHTMLタグによって論理構造を形成している．テーブル領域はテーブル記述用タグ(<TABLE>～</TABLE>)にてマークアップされ，プレーンテキストに比べHTMLソース内からの領域把握は容易である．また，テーブルは相互関係やコンテンツの要約・推移などに用いられ，保持する情報量は極めて高い．しかし，同時に<TABLE>タグは文書の調子を整えたり，レイアウト補正などにも用いられる為，<TABLE>タグのみでその領域がテーブルであるか否かを判断する事は出来ない．

WWW上のテーブル検出・分類については幾つかの研究がなされている[1],[2]．Chenら[1]は対象領域に特化したテーブル検出を行った．Wangら[2]は一般的な領域から広くテーブルを取得し，テーブル検出用データベースの構築を行っている．しかし，テーブルに含まれる情報の有用性を示唆した上でテーブル抽出を行う研究はなされているが，その抽出されたテーブルの具体的な活用手段については余り議論されていない．

筆者らは現在，WWW上の情報抽出・要約・統合研究の一環として，複数のパソコン(PC)の性能表を抽出し，各PCの特徴を抽出・比較する事によりユーザの要求に合致したPC選択を支援するシステムの構築を進めている[3],[4]．ここに，製品購入を目的とし，WWW上より情報収集を行おうとするユーザがいると仮定する．ユーザはYahoo[1]，Google[2]などの検索エンジンを利用して，もしくはメーカのウェブサイトにアクセスして製品情報を取得する．このプロセスを複数回繰り返し，幾つかの製品情報が手元に集まったとしても，収集した情報群の比較というさらに困難な作業が存在する．我々はこの様な情報群を有効に活用する事でもたらされる利便性を考慮し，本システムの開発を行っている．

システムの概要を図2に示す．まず，WWWより収集されたHTMLドキュメント群はフィルタリングによる選別を経て，性能表を含むドキュメントが性能表抽出処理へと渡される．性能表はメーカによって表記スタイルや項目名などの書式が異なる為，表構造と呼ばれるデータ構造へと正規化を行う[5]．得られた表構造中のデータを比較し，各製品の相対的な特徴にユーザの要求を反映したスコアリングを行い，スコアの序列に準じて，表の再構築，文章の生成，グラフ生成などの複数の形式を統合させた要約形式の製品仕様をユーザへと提供する．図3は我々が開発した製品選択支援システムである．

従来の研究において筆者らは，システムの処理対象を他種類の製品性能表へと拡張する際のコスト削減を狙い，性能表抽出への手掛かりとして用いるキーワー

[1]http://www.yahoo.co.jp/
[2]http://www.google.com/

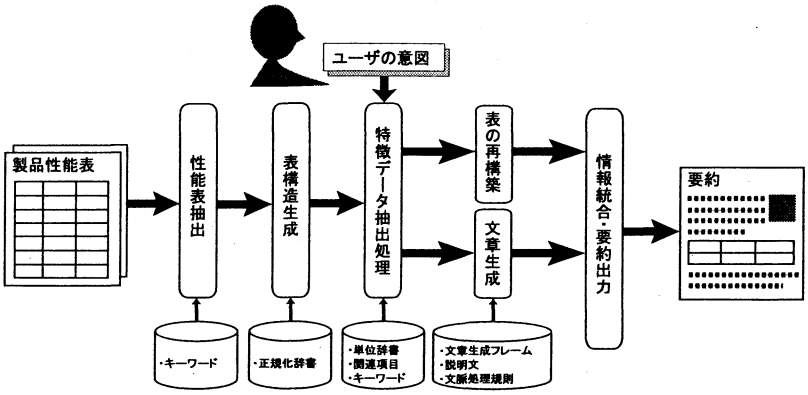


図 2: システム概略

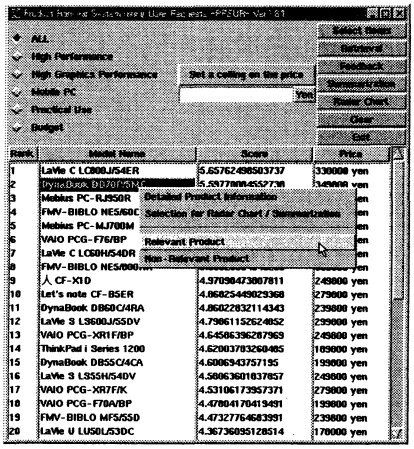


図 3: 製品選択支援システム

ドの自動生成法を提案した [6]. しかし，製品領域を PC に限定した場合においては手動によるキーワード生成とほぼ同一の精度を達成できたものの，実際にこの手法が他種製品へと拡張可能かといった評価は行っていなかった.

本稿では，このキーワード自動生成法の他種製品拡張への可能性について考察すると共に，製品領域に特化しない抽出規則の導入を目指す.

2 キーワードの自動生成

2.1 テーブルの氾濫

前述の通り，性能表は<TABLE>タグにて記述されている．性能表抽出とはこの<TABLE>タグでマークアップされた部分を抜き出す作業である．しかし，ある領域に特化したドキュメント群においては，ドキュメント 1 件あたり 2.35 個の<TABLE>タグが存在し，実際にテーブルとして用いられているものは 3 割に満たないという報告もある [1]. この事から，フィルタリングによってドキュメント内の性能表の有無を推定しなければならない．単一ドキュメント内に複数存在するテーブルの中から性能表を抽出する必要がある.

本研究では単語の頻度情報を特徴量とし，抽出に用いるキーワードを生成している．フィルタリング及び性能表抽出の詳細は 3 節にて述べる.

2.2 キーワードの定義

テーブル内に生起する単語の中でも，性能表の項目名は非常に特徴的な情報となり得る．よって，性能表抽出に必要なキーワードは

- 項目欄に位置する，テーブルの最左列
- 一定長以内の文章中

おいて顕著，または限定的に出現すると定義される．この定義に基づき，学習用ドキュメント群 D よりキーワード候補の抽出を行う．定義を候補抽出の条件として扱い，D においてこの条件を満たす文字列に対し形態素解析を行い，キーワード候補を抽出する．形態素解析には奈良先端科学技術大学で開発された「茶筌」[7] を用いた．次に，得られた候補に対して重みを加味する．情報理論の観点から語の特定性を考えれば，ある単語がドキュメント集合内の各ドキュメントにどれだけ偏って出現するかはエントロピーの概念より数値化する事が出来る．その数値を基にキーワード候補からキーワードを抽出する．ドキュメント d における単語 t の出現頻度を $tf(t,d)$ とすれば，各語の重み w_t^d は以下の式にて算出される.

$$w_t^d = log_2 \sum_{k=1}^{N} tf(t,k) + \sum_{i=1}^{N} \frac{tf(t,i)}{\sum_{j=1}^{N} tf(t,j)} log_2 \frac{tf(t,i)}{\sum_{j=1}^{N} tf(t,j)}$$

ここで，Dを性能表が存在するドキュメント群D_r及びそうでないドキュメント群D_nに分割し，各々について上述の重み付けを適用する．D_rに属するd_r及びD_nに属するd_nにおける単語tの重みをそれぞれ$w_t^{d_r}$及び$w_t^{d_n}$とすれば，ドキュメントdにおける単語tの重みw_t^dは下式へと拡張される．

$$w_t^d = \frac{w_t^{d_r}}{w_t^{d_n}}$$

この重みについて単語毎の総和$(ws_t^r = \sum_{i=1}^{N} w_t^i)$を取り，上位M位の単語をキーワードとみなす．上式の逆数の総和$(ws_t^n = \sum_{i=1}^{N} w_t^i)$も同様に求め，上位L位の単語をノイズワード，つまりドキュメント内に性能表が含まれない尺度として用いる．フィルタリングおよび性能表抽出にはこれらのキーワードおよびノイズワードを参照して判断する．

3 抽出処理

3.1 フィルタリング

ここでは前節にて自動生成されたキーワード及びノイズワードを用いて，ドキュメント内に性能表が存在するか否かの判定を行う．判定のルールは以下の通りである．

ドキュメントdにおけるキーワードの定義を満たす単語tについて，

$$r_r = \frac{\sum_{t \in K} w_t}{\sum_{i=1}^{M} w_{k_i}} \quad , \quad r_n = \frac{\sum_{t \in N} w_t}{\sum_{i=1}^{L} w_{n_i}}$$

$(K = |k_1, k_2 ... k_M|, N = |n_1, n_2 ... n_L|)$

を求め，

$$r = \begin{cases} r_r \times \frac{r_r}{r_n} & (r_n \neq 0) \\ r_r \times \frac{r_r}{0.01} & (r_n = 0) \end{cases}$$

が閾値を超えれば，ドキュメント内に性能表が存在するとみなす．閾値は任意の値を持つが，フィルタリングの性質上，再現率を優先した値を選択する．

3.2 性能表抽出

性能表抽出は以下の3つのプロセスにて構成される．

(1) テーブル内にキーワードが存在するか

(2) キーワード間最短テーブルの検索，仮取得

(3) 仮取得テーブルの閾値判定

段階を追って具体的な処理内容を説明する．

(1) ドキュメントのHTMLソースを走査的に解析し，<TABLE>タグ以降にキーワードの定義を満たす単語が存在するか否かを調べる．存在すれば単語をキーワードとして(2)へ遷移，存在しなければ終了する．

(2) 決定されたキーワードを含み，かつキーワードへの距離が最短のテーブルを探し，仮取得する．

(2.1) テーブルは入れ子構造を成す場合があるので，基本的にキーワードに対し最も近い<TABLE>タグが候補として挙げられる．最も近いテーブルが選出されれば(2.2)へ遷移する．

(2.2) テーブルがキーワードを含んで閉じて(</TABLE>タグ)いればテーブルを仮取得し，(3)へ遷移．テーブルが条件を満たさない場合はその次にキーワードに近いテーブルについて(2.2)へ再帰する．

(3) 仮取得したテーブル領域に存在するキーワードの重みの総和$sum = \sum w_t \quad (t \in |k_1, k_2 ... k_M|)$が閾値を越えれば性能表であるとみなして抽出，閾値以下であればテーブルを破棄し，(1)へと戻り処理を繰り返す．

処理(3)にて用いる閾値の選択については，フィルタリングにてある程度のノイズを除去出来たと仮定した上で，かつ，次行程の表構造生成モジュールへ渡すデータの高い信頼性を獲得する為，ここでは適合率向上を狙った閾値選択を考慮する．

4 実験と考察

以下においてフィルタリング及び性能表抽出の評価実験を行い，その結果について考察する．今回扱う製品種類はPC，デジタルカメラ(デジカメ)，DVD，プリンタの4種類とした．これら各製品ドキュメント群において，まずは学習データを抽出対象の製品性能表を含むもの，及び含まないものに分類し，それぞれL_r，L_n件ずつ用意した．L_r及びL_nの数は製品種類によって異なる．この学習データから，キーワード自動生成法によってキーワード30個，及びノイズワード15個を選出し，これらを参照してフィルタリング及び性能表抽出を行った．

まずはフィルタリングの実験内容について説明する．今回，3.1節にて述べたフィルタリングの規則r_r，r_n及びrを考案するにあたり，PC，デジカメ，DVDの3種類の製品ドキュメント群を用いた．これらの規則の正当性を評価する為に，プリンタ製品のドキュメント群を用いた．正解，ノイズはそれぞれT_r，T_n件ずつ用意した．精度評価には，情報検索分野において一般的な評価尺度である再現率，適合率及びF値を用いた．

$$\text{再現率}\ (R) = \frac{\text{正しく抽出されたドキュメント数}}{\text{性能表を含むドキュメント数}} \times 100$$

$$\text{適合率}\ (P) = \frac{\text{正しく抽出されたドキュメント数}}{\text{抽出された全てのドキュメント数}} \times 100$$

$$F\ \text{値}\ (F) = \frac{1}{\alpha \frac{1}{R} + (1-\alpha)\frac{1}{P}}$$

F値の式におけるαはR及びPの相対的な重みを表す．フィルタリングでは再現率を優先し，α=0.6とした．表1はデータの内訳，また表2は閾値によるフィ

表 2: 閾値に伴うフィルタリングの精度変化

	PC			デジカメ			DVD			プリンタ		
閾値	R	P	F	R	P	F	R	P	F	R	P	F
0.15	100	79.4	90.58	100	58.5	77.90	100	60.0	78.95	100	73.7	87.50
0.20	100	80.7	91.24	100	63.7	81.44	100	61.5	80.00	99.0	76.4	88.50*
0.25	100	81.3	91.58*	100	65.0	82.30	100	61.5	80.00	94.9	79.5	88.07
0.30	95.0	82.6	89.62	100	66.0	82.89	100	64.9	82.19	91.8	81.8	87.55
0.35	92.0	82.1	87.79	100	67.4	83.78	100	64.9	82.19	87.8	81.9	85.32
0.40	91.0	84.3	88.18	100	68.4	84.39	100	66.7	83.33*	85.7	82.4	84.34
0.45	90.0	85.7	88.24	100	68.9	84.70	100	66.7	83.33*	84.7	84.7	84.69
0.50	89.0	86.4	87.94	98.9	68.7	84.10	87.5	65.6	77.21	81.6	87.0	83.68
0.55	88.0	88.9	88.35	98.9	70.2	85.03	87.5	65.6	77.21	79.6	86.7	82.28
0.60	88.0	88.9	88.35	97.9	72.8	86.01	87.5	67.7	78.36	77.5	87.2	80.47
0.65	86.0	89.6	87.40	96.8	73.2	85.71	87.5	75.0	82.03	76.5	87.2	80.47
0.70	85.0	89.5	86.73	96.8	73.8	86.04	87.5	75.0	82.03	76.5	87.2	80.47
0.75	84.0	89.4	86.07	96.8	73.8	86.04	83.3	76.9	80.65	73.5	86.8	78.26
0.80	82.0	90.1	85.06	94.6	75.9	86.11*	83.3	80.0	81.97	73.5	86.8	78.26
0.85	79.0	91.9	83.69	91.4	76.6	84.83	83.3	80.0	81.97	73.5	86.8	78.26
0.90	79.0	91.9	83.69	90.3	77.1	84.51	83.3	80.0	81.97	73.5	87.8	78.60
0.95	79.0	91.9	83.69	89.3	78.3	84.52	83.3	80.0	81.97	68.4	87.0	74.78

*印はその製品における最大 F 値

ルタリングの精度の変化を示したものである.

次に, フィルタリングの実験結果を用いて性能表抽出を行う. ここでは, フィルタリング行程にて最良の精度を持つ結果 (プリンタ, 閾値=0.20, F=88.50%) を用いた. 表抽出の閾値は $(\sum_{t \in K} w_t)/2$ とし, F 値は適合率を優先し, α=0.4 にて評価した. 結果を表 3 に示す.

実験結果について考察する. これまで PC 製品のみを対象としていた本手法を他種類製品に拡張した場合においても 80%程度の精度が達成された. この事から, 本稿にて提案したキーワード自動生成法の有用性が確認された. しかし, 各製品種類を比較してみると, 製品によって最良の精度を持つ閾値の差が激しい事が分かる. これについては, フィルタリングの規則が正解及びノイズ各々の特徴を把握しきれていないと推測され, フィルタリング規則の再検討が求められる. また, 性能表抽出処理におけるノイズ誤認及び抽出洩れについては, 様々なテーブルの書式に対応した表抽出アルゴリズムの具体化が今後の課題となる.

表 1: データの内訳

製品種類	L_r : L_n	T_r : T_n
PC	50 : 50	100 : 100
デジカメ	50 : 50	93 : 100
DVD	12 : 50	24 : 100
プリンタ	50 : 50	98 : 100

表 3: 実験結果

	正解 (98)	ノイズ (100)	R	P	F
フィルタリング	97	30	99.0	76.4	88.50
性能表抽出	72	10	74.2	87.8	81.80

5 おわりに

本稿では, これまで PC 製品のみを処理対象としていた性能表抽出処理の他種類製品への拡張を目指し, 領域依存しない普遍的な特徴からのキーワード生成及びフィルタリング規則の考案について述べた. 抽出処理に用いるキーワードはエントロピーによる重み付け学習から自動生成し, キーワードの重みを考慮した規則によりフィルタリングを行った. これらの手法を用いた評価結果は 80%の精度を達成した.

今後はフィルタリング規則の改良を展望とすると共に, 表抽出処理の最適化, 及び本手法とは異なるアプローチとの比較考察なども研究視野に入れている.

参考文献

[1] H. H. Chen, S. C. Tsai and J. H. Tsai: Mining tables from large scale HTML texts, Proc. of the COLING2000, pp.166-172, 2000.

[2] Y. Wang and J. Hu: A machine learning based approach for table detection on the Web, Proc. of The Eleventh International World Web Conference, 2002.

[3] A.Fukumoto et.al.: Information Extraction from Specifications on the World Wide Web, Proc. of the PACLING2001,pp.109-116,2001.9

[4] 森松俊允, 福本篤史, 嶋田和孝, 遠藤 勉: 製品性能表を用いた製品選択支援システムの構築, 電気関係学会九州支部連合大会, 2002.

[5] 福本篤史, 嶋田和孝, 遠藤 勉: 製品性能表からの表構造生成, 電気関係学会九州支部連合大会, 2002.

[6] 林 晃司, 嶋田和孝, 遠藤 勉: Web ページからの製品性能表抽出, 第 10 回電子情報通信学会九州支部学生会, JCEEE, 2002.

[7] 松本裕治, 北内啓, 山下達雄, 平野喜隆, 松田寛, 高岡一馬, 浅原正幸: 日本語形態素解析システム「茶筌」 http://chasen.aist-nara.ac.jp/