

文脈を用いたメディア間の対応付けと
情報抽出に関する研究

大分大学大学院工学研究科

博士後期課程

博士論文

2003年3月

嶋田 和孝

もくじ

1	序論	1
1.1	研究背景と目的	1
1.2	論文の構成	4
2	メディア情報と文脈	5
3	対話からの情報抽出	8
3.1	はじめに	8
3.2	基本的な考え方	9
3.2.1	問題解決	10
3.2.2	教師との対話	10
3.2.3	対話における文脈	12
3.2.4	システムの設計方針	14
3.3	システム構成と文脈表現	15
3.3.1	言語情報の意味表現	15
3.3.2	図形情報の意味表現	16
3.3.3	コマンドの表現	18
3.3.4	文脈の表現	19
3.4	問題解決過程と文脈処理	21
3.4.1	文脈情報の初期値設定	21
3.4.2	プランニングとコマンド付加手続き	21
3.4.3	発話文の生成	22
3.4.4	応答文の解釈	23
3.5	実験と考察	24
3.5.1	対話例	24
3.5.2	対話処理部の評価	30

3.5.3	試作システムの問題解決能力	31
3.5.4	対話支援型問題解決システムの意義	32
3.5.5	言語表現と図形表現の統合化	33
3.6	まとめ	33
4	表からの情報抽出	35
4.1	はじめに	35
4.2	関連研究	38
4.3	製品性能表の抽出	40
4.3.1	キーワード抽出	40
4.3.2	エントロピーによる重み付け	41
4.3.3	ベイズの定理による重み付け	42
4.3.4	性能表抽出処理	42
4.3.4.1	フィルタリング	42
4.3.4.2	表領域抽出	43
4.3.5	実験と考察	44
4.3.6	他研究との比較	47
4.4	製品性能表の解析	48
4.4.1	表構造の定義	49
4.4.2	表構造の生成	49
4.4.3	実験と考察	51
4.5	特徴データ抽出処理	53
4.5.1	定量的項目の特徴データ抽出	53
4.5.2	定性的項目の特徴データ抽出	55
4.5.3	ユーザの要求に応じたスコアリング	57
4.5.4	適合性フィードバック	59
4.5.5	実験と考察	59
4.5.5.1	特徴データ抽出処理実験	59
4.5.5.2	ユーザの要求と適合性フィードバックを用いた検索	63
4.6	出力形式	65
4.6.1	文章による要約生成	65
4.6.1.1	文章生成フレームと説明文	66
4.6.1.2	文章生成処理	67

4.6.1.3	修辭処理	68
4.6.2	表の再構成	69
4.6.3	レーダチャートの生成	70
4.6.4	出力結果に対する評価	71
4.7	システム全体に対する考察	73
4.7.1	他種類データへの移行	73
4.7.1.1	表抽出処理の移行	73
4.7.1.2	特徴データ抽出処理の移行	75
4.7.2	関連研究との比較	78
4.7.3	対話システムと文脈処理	79
4.8	おわりに	80
5	文章および画像からの情報抽出	82
5.1	はじめに	82
5.2	画像と隣接文章の抽出	83
5.2.1	対象画像の抽出	83
5.2.2	隣接する文章の抽出	85
5.3	キーワードと重みづけ	86
5.3.1	人手によるキーワードと重み	86
5.3.2	<i>tf · idf</i> 法による重み	87
5.3.3	ベイズの定理による重み	88
5.4	評価実験	89
5.4.1	画像抽出処理	89
5.4.2	画像分類処理	89
5.5	おわりに	92
6	結論	93
	謝辞	96
	参考文献	97

表一覧

3.1	ドリルテキストの問題解決・解答合成用のコマンド	12
3.2	問題解決過程の対話における発話意図とその発端となる状況	14
4.1	エントロピー $M_{real} = 30, M_{no} = 15$	44
4.2	エントロピー $M_{real} = 56, M_{no} = 37$	45
4.3	ベイズの定理 $M_{real} = 19, M_{no} = 15$	46
4.4	ベイズの定理 $M_{real} = 66, M_{no} = 41$	47
4.5	表領域抽出処理 - エントロピー -	47
4.6	表領域抽出処理 - ベイズの定理 -	48
4.7	表構造生成処理の実験結果	53
4.8	定量的項目と定性的項目	54
4.9	比較のための単位の分類	54
4.10	標準値の算出処理のための分類	55
4.11	ユーザからの要求と関連する属性	57
4.12	属性間の関連性	58
4.13	抽出された特徴データ	62
4.14	実験結果	63
4.15	システムによって1位にランクされた製品の雑誌中での順位	64
4.16	雑誌中で1位にランクされた製品のシステム中での順位	64
4.17	話題と対応する属性	68
4.18	出力形式による読む側のわかりやすさ	71
4.19	生成された文章の評価	73
4.20	エントロピー $M_{real} = 30, M_{no} = 15$ - デジタルカメラ -	75
4.21	ベイズの定理 $M_{real} = 29, M_{no} = 23$ - デジタルカメラ -	75
4.22	表領域抽出処理 - エントロピー・デジタルカメラ -	76
4.23	表領域抽出処理 - ベイズの定理・デジタルカメラ -	76

4.24	比較のための単位の分類 - 携帯電話 -	77
4.25	標準値の算出処理のための分類 - 携帯電話 -	77
4.26	比較のための単位の分類 - デジタルカメラ -	77
4.27	標準値の算出処理のための分類 - デジタルカメラ -	77
5.1	タグの重み	86
5.2	対象画像抽出処理の結果	89
5.3	画像分類処理の結果 - 訓練データ -	91
5.4	画像分類処理の結果 - テストデータ -	91

図一覧

2.1	本論文で扱う文脈処理の対象	7
3.1	小学校1年算数のドリルテキストの例	11
3.2	問題解決過程における対話例	13
3.3	対話支援型問題解決システム	15
3.4	ドリルテキストの意味表現	17
3.5	クラスコマンドフレームの例	19
3.6	コマンドフレームの生成過程	26
3.6	コマンドフレームの生成過程(続き)	27
3.7	システムと教師の対話例	28
3.8	文脈情報の例	29
4.1	製品性能表の例	37
4.2	システムの概要	37
4.3	製品選択支援システム	38
4.4	性能表抽出の流れ	41
4.5	HTML表記による表の例	50
4.6	対応付けられた表	50
4.7	行項目の特定	52
4.8	表構造の例	52
4.9	スコアリング	55
4.10	関連項目による特徴データ抽出の例	56
4.11	重み選択ウィンドウ	58
4.12	実験に用いた製品性能表の例	61
4.13	適合率と再現率	65
4.14	文書の構造	66
4.15	文章生成処理の例	69

4.16	レーダチャート	70
4.17	生成された文章の例	72
4.18	表と文章を統合した出力	74
4.19	製品選択支援システムにおける対話例	80
5.1	製品紹介サイト	84
5.2	画像のタイプ	84
5.3	レイアウトとその HTML 表記	85
5.4	キーワード抽出処理の例	88
5.5	画像とそれに隣接する文章	90
5.6	キーワード数と精度	92

第 1 章

序論

1.1 研究背景と目的

文脈とは第一義的には文の脈絡のことを意味し，文と文あるいは発話と発話間の論理的な関係を指すものとして用いられてきた．従って，文脈処理は自然言語処理研究の重要な研究課題として扱われることが多く，これまでに，

- (1) 文章の構造理解
- (2) 照応処理
- (3) 概念や事象間の関係理解

などについて研究が行われてきた [38]．(1) では，文章のおおまかな流れを示す「主題」や話し手の協調したい「焦点」の解析，等位接続や選言接続など文間の結束構造の特定処理などが行われる．(2) では，ある言語表現が，後に現れる言語表現と同じ内容や対象を指すときに生じる照応関係における先行詞と照応詞の同定，文章生成における代名詞化などが扱われる．(3) では，文間における時間的变化，空間的变化，状態変化などの事象系列の理解やそのための因果関係知識の作成などが行われる．

一方，文脈を広義に解釈し，言語表現や発話の背景，即物的な場に関わる具体的な事物や対象まで含めて考えることもある [80]．これを言語的な文脈に対して，言語外的な文脈と呼ぶ．近年，この言語外的な文脈の重要性が急速に高まってきている．その例を対話とマルチメディア文書の場合で見よう．前者の対話においては，一般に，生成される発話文や事前に行われた発話内容，発話意図，現在の発話がどの発話に対応するのかといった情報を言語的な文脈として保持しておき，この文脈を適宜参照することで，発話の冗長性解消や曖昧性解消，自らが発話する際の語の省略や代名詞化，相手の発話中で省略さ

れた語や代名詞の指す内容の理解などが可能になる．ところが，何らかの問題を解決しようとする際に行われる対話の理解や生成においては，言語的な文脈に加えて，言語外的な文脈が重要になってくる．例えば，言語以外の手段で表現された解決すべき問題，問題解決を行う環境，問題解決過程の履歴などである．これらの情報と対話文の対応付けを考えなければ，対話の理解や生成は不可能といえる．さらに，言語発話だけでなく，視線や表情，身振りや手振りなどのジェスチャを併用したマルチモーダル対話においては，言語外的な文脈の役割がより高いものになってくる．

後者のマルチメディア文書の典型的な例が Web 上の文書である．この文書には，自然言語による文章だけでなく，表やグラフ，図形や画像などが用いられている．検索や要約といった Web 文書の計算機処理では，文章のみを対象にして行われることが多い．しかしながら，表，グラフ，図形，画像には文章では表現できない情報が含まれており，相互に関係をもちながら文書が構成されている．従って，文章だけを解析対象としたのでは，十分な検索結果や要約が得られるとは限らない．例えば，あるユーザがある製品を購入しようとした場合を考える．一般に，そのユーザは関連する製品の特徴などを記述したレビューなどの文章だけでなく，その製品の詳細な機能や性能などを記述した表を見比べることで，自分の欲しい製品を探す．外観に重きを置いている場合は，画像などによってその製品をチェックする．Web 文書においては，文章に対して，表，グラフ，図形，画像などが言語外的な文脈に相当する．

しかしながら，言語外的な文脈を取り扱う確立した統一的な手法が存在するわけではなく，これがどのような構造をしているのか，文書や対話の中でどのような機能を果たしているのかといった問題は今後解決すべき重要な課題として残されている．

本論文では，文脈をメディア内ならびにメディア間の脈絡として捉え，この文脈を用いたメディア間の対応付けおよび情報抽出への応用について述べる．具体的には，言語外文脈が，その解析や生成に大きく寄与すると思われる問題解決向き対話とマルチメディア文書を対象にして

- (A) 対話からの情報抽出
- (B) 表からの情報抽出
- (C) 文章および画像からの情報抽出

という 3 つの課題について議論する．(B) の表と (C) の文章および画像はマルチメディア文書中に現れるものである．

(A) では，問題解決過程における対話と文脈処理について考察する．問題解決の対象は小学校 1 年算数のドリルテキストである．発話の環境は「生徒（計算機）が教師（ユー

ザ)と対話をしながら,ドリルテキストの意味を理解し,その問題を解く」という状況に設定している.計算機はドリルテキストから問題解決に必要な情報を取り出し,その解決を試みる.この過程で不具合が生じると,教師との会話を通して更なる情報抽出を図る.この場合の文脈とは,言語表現が主体の対話であることから,「言語表現が使用され,評価される言語的・非言語的環境である」と考えている.具体的には,(1)先行発話文とその発話意図,(2)発話のおかれた状況(話し手,聞き手,場所,時,現場の事物など),(3)領域知識(発話の状況に依存する),(4)話し手の注意状態(話題や焦点など)で構成される.さらに,文脈に基づく対話制御方式を提案する.ここで,文脈に基づく対話制御方式とは,計算機からの発話文の生成においては,問題解決過程の状況からの発話意図の発生と言語情報の抽出法や文脈情報を利用した修辞処理,教師からの応答文の解析においては,発話意図や省略要素の推定,発話中の言語表現と文脈情報との対応付けなどを指す.

(B)では,Web 文書中の製品性能表からの情報抽出と統合ならびに要約処理について考察する.まず,文章とは表現形式の異なる表の解析およびその表から抽出された情報の活用法について論じる.表データを用いることの意義と有効性を示すために,パーソナルコンピュータ(PC)を対象とした製品選択支援システムの構築を行う.形式の異なる記号メディアで表現された情報を統合する意義として,文章や表,グラフを用いることで,ユーザにより分かりやすい情報提示を行うことの有効性も検証する.この場合の文脈とは,表現形式の異なるメディアを扱うことから,「異種形式メディアが混在する環境から各メディアを同定して相互に関連付け,同一形式メディア内の情報を標準化する,または他形式へ変換するための特徴である」と考えている.具体的には,(1)表を同定するためのキーワードとその重み,(2)正規化規則,(3)特徴抽出規則,(4)形式変換のためのフレーム(本論文では文章生成フレームなど),(5)修辞処理規則などから構成される.

(C)では,Web 文書中の画像の意味内容を画像に隣接した文章から抽出する手法について考察する.Web 文書中の画像は,一般に低解像度であり,従来の画像解析手法を用いた画像内容推定は必ずしも適さない.しかしながら,画像とこれに隣接する文章とは密接な関係がある.そこで,画像の内容を隣接する文章情報を用いて推定する手法を提案する.この場合の文脈とは,画像と文章という異種メディアを扱うことから,「2つのメディアが混在する環境から各メディアを同定して相互に関連付け,一方のメディアから他方のメディアの意味内容を推定する特徴である」と考えている.具体的には,(1)画像の空間的特徴,(2)画像と文章の空間的位置関係とHTML タグ情報,(3)画像内容を推定するための文章中のキーワードとその重みなどから構成される.

1.2 論文の構成

本論文は、本章を含めて全 6 章からなっている。以下に各章の概要を示す。

2 章では、各種メディア情報と関連付けた文脈の定義について考察し、本論文で提案する 2 つのシステム（対話支援型問題解決システム、製品選択支援システム）との関連性を明らかにする。

3 章では、小学校 1 年算数のドリルテキストを対象とした問題解決システムについて述べる。システムは対話などから情報を抽出し、それらから構成される文脈情報を用いて問題解決を図る。システムを構築する際の基本的考え方、文脈情報の表現法を示し、これに基づく発話文生成と応答文解釈の方式を提案する。Prolog 上に対話支援型問題解決システムを構築し、対話処理の評価、本システムの位置付けおよび意義について議論する。

4 章では、表からの情報抽出手法ならびに表からの文章生成手法について提案し、その応用例として、表から抽出されたデータを用いた製品選択支援システムについて述べる。まず、Web 上からの表の抽出と 2 次元である表を 1 次元の表構造と呼ばれるリスト構造に変換する手法を述べる。得られた複数の製品の表構造を基に各製品の相対的な特徴を抽出し、スコアリングする手法について述べ、その結果をどのような形式で出力することがユーザにとって分かりやすいかを考察する。作成された製品選択支援システムのプロトタイプを用いて、表の抽出・解析、特徴データ抽出処理、スコアリング、出力形式、汎用性などの観点から実験・考察を行い、その有効性を検証する。

5 章では、4 章で提案した製品選択支援システムの出力形式である表と文章の要約に画像を統合し、より分かりやすい要約出力・情報提示を実現するための画像内容の分類手法について述べる。画像データの詳細な解析は行わずに、画像に隣接する文章から、その画像の内容を表すキーワードを抽出し、重み付けを行う。これらのキーワード群を用いて画像の内容推定を行う手法を提案する。この手法により、画像の解像度などに左右されないシステムの構築が可能となる。キーワードとして (1) 人手によるもの、(2) $tf \cdot idf$ 法によるもの、ベイズの定理によるもの、の 3 種類を定義し、精度比較を行う。実験結果から、文章・画像間の文脈情報を利用した本手法により、優れた画像分類精度が得られることを示す。

6 章では、本研究を総括し、今後の課題について述べる。

第 2 章

メディア情報と文脈

メディアという言葉は多様な意味に用いられる [81]。本論文では、コンテンツ内容を表現する要素となるメディアの意味で使用する。このメディアは記号化された記号メディアと記号化されていないパターンメディア（非記号メディア）に大きく分けられる。記号メディアは、論理的性質に重点を置いて対象を表現する。自然言語と呼ばれる言語、数式やプログラム言語、知識表現言語などの形式言語、グラフや表、音声言語などである。パターンメディアは、物理的性質に重点を置いて対象を表現する。映像、画像、音響、表情や視線、身振りや手振りなどである。それぞれのメディアにはそれぞれの役割があり、他のものでは代替することができない役割をもっている。さらに、これらのメディアは単独で用いられることは少なく、テレビや新聞などからも分かるように複数のメディアが併用されている。特に、最近の計算機を取り巻く環境は急速にマルチメディア化、マルチモーダル化が進んでいる。

本論文では対話支援型問題解決システムおよび製品選択支援システムという 2 種類のメディア情報処理システムを扱っており、各システムの有用性、特徴、関連システムとの比較評価などについては、第 3～5 章で詳述する。本章では各システムで対象とするメディア情報の種類と文脈情報との関連性について考察する。

まず、対話支援型問題解決システムでは日本語による文字ベースの対話文とドリルテキストの問題文という典型的な記号メディアに加えて、問題図形というパターンメディアを扱う。ただし、問題図形はシステム内部では図形構造表現という記号メディアで表現されている。現在、対話のモダリティとして音声言語と指示や描画などのジェスチャを加えたマルチモーダル対話システムへの拡張を進めている。一方、製品選択支援システムでは、情報源としてインターネット上の Web 文書を扱っており、文章、表、グラフという記号メディアと画像というパターンメディアを考える。

両システムにおける各メディアは一般的には相補的な関係にあるものの、メディア間

の時間的・空間的關係，キーワードや概念（意味素性）を介在させた連携など相互関係をもっており，あるメディアの解析には他のメディアの情報が不可欠といえる．そこで，本論文では，あるメディアが使用され，評価されるとき他のメディアやメディア間の関係を広い意味での文脈と捉えることにする．すなわち，本論文における文脈とは，メディア内ならびにメディア間の脈絡として定義する．この脈絡が文脈情報となるが，その表現形式は多様である．以下に文脈情報の種類と表現方法の概要を示す．

(1) 複数のメディアが混在する中でのメディア間の脈絡

- 文書内でのメディア間の空間的位置関係：HTML タグ情報など
- 各メディアの内容的特徴：メディアを特徴付けるキーワード群とその重みなど
- メディア間で共起する概念：意味素性など
- メディア間を関連付けるメタ的枠組み：形式変換のためのフレームなど
- メディアを使用するときの意図：問題解決過程の履歴など
- 領域知識（対象とする対話や文書に依存する）

(2) 同一メディア内での要素間の脈絡

[言語メディア]

- 文章における文間の空間的位置関係
- 対話文における時間的前後関係：先行発話文の履歴，発話対など
- 話し手や書き手の意図：発話意図，注意状態（話題や焦点など）
- 文間あるいは発話間の連結性：省略，照応，提題などの修辞処理など
- 文間あるいは発話間の首尾一貫性：文章内でのキーワードの共起など

[表メディア]

- 表の構造的特徴：HTML タグ情報など
- 表の形式と内容との関係：正規化規則など

[画像メディア]

- 画像内の対象物あるいは背景との空間的位置関係
- 画像データの特徴：サイズ，濃淡，色，テクスチャ情報など

これらの文脈情報と本論文で提案するシステムとの関連を図 2.1 に示す．

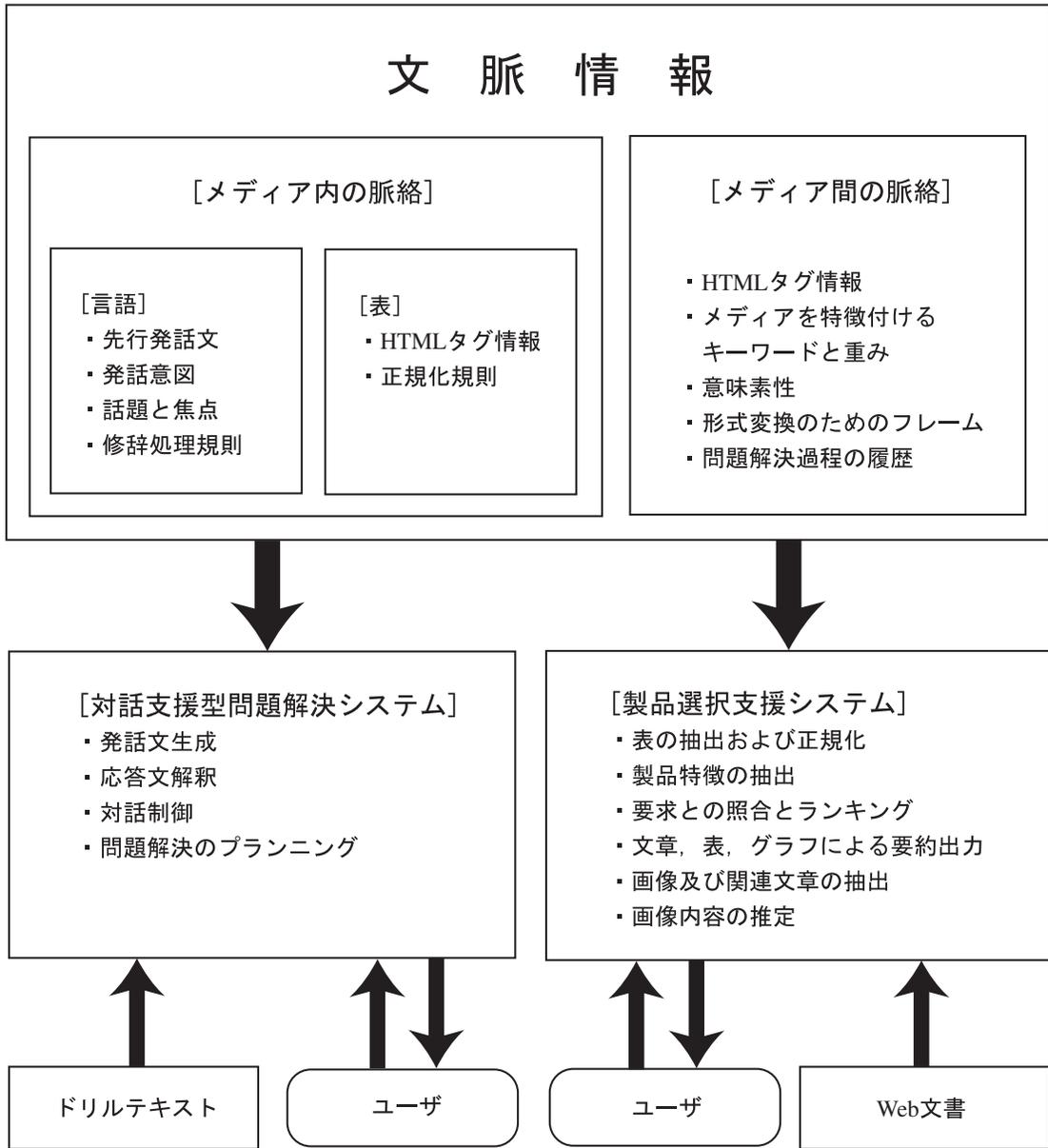


図 2.1: 本論文で扱う文脈処理の対象

第 3 章

対話からの情報抽出

「生徒（計算機）が教師（ユーザ）と対話をしながら，ドリルテキストの意味を理解し，その問題を解く」という状況での，対話型問題解決システムについて述べる．システムは，文字ベースの対話文とドリルテキストの問題文という典型的な記号メディアに加えて，問題図形というパターンメディアを扱う．このシステムは現在，入力および出力に音声とジェスチャを用いたマルチモーダル化を進めているが，本章では，特に問題解決過程における対話からの発話意図や焦点要素などの情報抽出およびそれらを用いた文脈処理について述べる．

3.1 はじめに

問題解決能力の習得や知識の形成において対話が重要な役割を果たすことは論を待たない．初等教育においても，話し言葉中心でなされる生徒の日常生活からの自然な延長として，教師と生徒との対話が教育活動の主要部分を占めており，教師の発話を生徒が容易に理解できるように教科書やドリルなどの教材，黒板に書かれた文章や図表を適宜参照しながら対話を進めるのが普通である [72]．しかも，対話と教材との対応が円滑に行えるように，教材の文章は話し言葉に近い形式で書かれている．逆に，教材の参照が文法からの逸脱，省略語や指示語の頻繁な出現という，書き言葉と違った話し言葉固有の問題点を解決する一助になっているともいえる．このような対話と各種教材の相互参照に基づく問題解決方式を計算機上で実現しようとするとき，教材ならびに対話文の意味表現と問題解決のための知識表現との整合，問題解決過程での適切な発話文の生成と応答文の解釈，対話および問題解決過程の履歴を保存した文脈情報の表現，複数表現の協調による曖昧性解消と照応処理などの機能が要求される．

筆者らは，言語表現と図形表現が相補的に機能する形で構成されている小学校 1 年算数

のドリルテキストの世界 [52][53] を対象として、教師（人間）との対話とテキストを相互参照しながら算数の問題解決と知識獲得を行うシステムの開発を進めている。

算数の問題解決については、教師と対話をしながら知識を獲得するシステム LAPS があるが [51]、問題が文章題に限られており、また対話のための文脈処理も考慮されていない。岩根らのシステムも算数の加減算の文章題を扱っているが [24]、目的が生徒の学習支援であり、システムからの質問は生徒の理解過程を確認するためのものである。説明にはアニメーションも使われているが、問題自体に図形表現が含まれているわけではない。英文と図形で表現された物理学の問題を解決しようとする Novak らのシステム [42] や図を用いて算術問題の解決を行う村田らの DIPS [37] は、筆者らの狙いに近いものであるが、対話処理との統合は想定されていない。

一方、対話文の理解や生成については従来多くの方式やシステムが提案されているが [22][30][46][78]、対話の主題が情報検索や質問応答などであり、ユーザである人間の質問を解析して発話意図を認識し、協調的な応答文を生成しようとする枠組みが大部分である。システム主導の対話、すなわち問題解決や知識獲得過程のどのような時点で、どのような発話文を生成するかといった側面からの研究はまだ十分になされていない。さらに、文脈処理については、対話文理解を目的とした文脈情報の定式化の試み [66]、文脈情報を利用した自然言語文の構造的曖昧性の解消 [39] などが提案されているが、発話文生成過程での利用や対話の現場にある事物との関係付けなどは検討されていない。

本章では、教師と対話をしながら算数のドリルテキストの問題を解決するシステムを構築する際の基本的な考え方を示し、これに基づく発話文生成と文脈処理方式を提案する。以下、3.2節で対話支援型問題解決に対する基本的考え方を示し、3.3節でシステム構成と文脈情報の表現法について述べる。3.4節で文脈情報に基づく発話文生成と応答文解釈のアルゴリズムを提案し、3.5節で実験システムによる問題解決と対話処理の評価ならびに考察を行う。

3.2 基本的な考え方

対象としている小学校1年算数のドリルテキストの例を図3.1に示す。テキスト中には純粋な文章題も含まれるが、多くは言語表現と図形表現を併用した形で問題が与えられている。ここでは、教師と対話をしながらこの種の問題解決を計算機に行わせようとする際の基本的な考え方について述べる。

3.2.1 問題解決

小学校1年算数の分野は、数の意味（なかまあつめ，1対1対応，集合数，順序数，無の意味），加減算（合併，添加，求残，求差，記号と式），計測（長さ比べ，かさ比べ），形，分け方，時計などから構成されている [72]．これらの内容を生徒が理解しているかどうかを確認するために作られたドリルテキストでは，様々な出題形式の工夫がなされている．加減算については純粋な文章題もあり，計算機による文章題の解決法式が提案されている [24]．しかしながら，多くの問題は文章と図形を相互に参照しながら解く必要があり，しかも参照方式が多様であるため，一連の統一的な形式で問題解決手続きを構成するのは難しい．

そこで，ドリルテキストの問題解決や解答合成に必要な要素的手続きを考え，出題形式に応じてこれらの手続きを組み合わせることで計算機に問題解決を行わせることにした．このような要素的手続きをコマンドと呼ぶことにする．2種類のドリルテキストに現れた260個の問題を分析し，30個のコマンドを抽出した．その一部を表3.1に示す．コマンドは，大きくオペレーション，述語，制御に分けられる．述語はオペレーションのパラメータの中で，条件を記述するとき使われる．表3.1の中で，count，order，addなどは，問題図形の認識により得られた記号表現に対して行われるオペレーションであるが，drawやmeasure-lなどは図形データに対して直接作用するものである．

3.2.2 教師との対話

計算機はドリルテキストの問題文から，その解決に必要なコマンド候補を抽出し，さらに問題図形から各コマンドのパラメータを決定した後，実行を試みる．この過程で不具合が生じると，教師との対話を通してその解決を図る．従って，対話は計算機主導で行われることになる．

そこで，実際にどのような対話が想定されるかを調べるために，本問題解決過程の枠組みに従って人手によりシミュレーションを行った．対話の進め方については文献 [72] を，省略・指示・提題など対話文の表現法については文献 [34] を参考にした．図3.2は図3.1の(a)のテキストを解くときに想定された対話例である． S_i は計算機からの， T_i は教師からの発話を示す．対話は基本的に S_i と T_i の発話対から構成されており，2つの発話間の関係は各々の発話の役割に応じて類型化できる [22][66]．対話の各発話対における発話の役割を発話意図と呼ぶことにする．シミュレーションにより得られた対話例を分析することにより，表3.2に示す発話意図を抽出した．これらの発話が問題解決過程のどのような状況でなされるかについても併せて示している．例えば，図3.2の S_1 と T_1 は情報要求

① かずだけ ○を かきましょう。

①					
②					
③					

(a)

① おなじ かずの かあどを —
 せて むすびましょう。

あ い う え

●	●	●	●
0	1	2	3

(b)

図 3.1: 小学校 1 年算数のドリルテキストの例

表 3.1: ドリルテキストの問題解決・解答合成用のコマンド

コマンド名	コマンドの機能	
オペレ	Count (List, Num)	List 中のオブジェクトの個数 (Num) を数える
	Order (List, S, N, Obj)	List 中の S から N 番目のオブジェクト (Obj) を求める
レ	Select (List, [C], Obj)	List 中から条件 C を満たすオブジェクト (Obj) を選ぶ
	Add (N1, N2, Sum)	数値 N1 と N2 の和を Sum とする (差 Sub も同様に定義する)
シ	Identify (Obj, [C])	条件 C を満たすオブジェクト (Obj) を図形表現中から同定する
	Write (Sym, Loc)	記号オブジェクト (Sym) をテキスト中の位置 (Loc) に書く
ヨ	Draw (Fig, Loc)	図形オブジェクト (Fig) をテキスト中の位置 (Loc) に描く
	Link (Obj1, Obj2)	2 つのオブジェクト Obj1 と Obj2 を線で結ぶ
ン	Measure-l (Obj, Len)	図形オブジェクト (Obj) の長さ (Len) を求める
	Eqn (Num1, Num2]	数値 Num1 と Num2 が等しい
述	Eqc (C1, C2)	概念 C1 と C2 が等しい
	Gt (Num1, Num2)	数値 Num1 が Num2 より大きい (小さいときは lt で表現)
話	Repeat (Num,	オペレーション系列 (Operation-seq) を Num 回繰り返す
	[Operation-seq])	(実行時には, Num 個のオペレーション系列に展開する)
制		
御		

と提供からなる発話対, S3 と T3 は確認要求と同意表示の発話対である。また, S2 は情報提供と情報要求を同時に行うための発話である。

3.2.3 対話における文脈

文脈の定義については諸説があるが, ここでは言語表現が使用され, 評価される言語的・非言語的環境であるという定義を採用する [68]。この環境の構成要素として, 文献 [66] で提案されている定式化も参考にし, 次の項目を考えた。

- (1) 先行発話文とその発話意図
- (2) 発話のおかれた状況 (話し手, 聞き手, 場所, 時, 現場の事物など)
- (3) 領域知識 (発話の状況に依存する)
- (4) 話し手の注意状態 (話題や焦点など)

発話の環境を「計算機が教師と対話をしながらドリルテキストの意味を理解し, その問題を解く」という状況に設定した場合の, 具体的な文脈の構成要素と表現法については 3.3.4 で述べる。

S1: なんの かずを かぞえるのですか。
T1: わくの なかの ものです。
S2: わくの なかには いろいろな ものが あります。
 なんの かずを かぞえるのですか。
T2: まず ちゅうりっぷの かずを かぞえなさい。
S3: これの かずだけ まるを かくのですね。
T3: そうです。
S4: どこに かくのですか。
T4: したの ひょうです。
S5: ひょうには くらんが たくさんあります。
 ひょうの どこに かくのですか。
T5: ひだりはしに ちゅうりっぷの ある ぎょうです。
S6: ぎょうの どこから かくのですか。
T6: 2れつめからです。
S7: これで よいのですか。

図 3.2: 問題解決過程における対話例

表 3.2: 問題解決過程の対話における発話意図とその発端となる状況

発話意図	発話者	発話の発端となる状況
確認要求: CONREQ (CONfirmation REQuest)	計算機	知識あるいは教師の発話文から新しいコマンド候補が生成できたとき コマンドの具体的なパラメータ値を推論手続きにより求めたとき 最終的なコマンド系列を実行して問題の解答を求めたとき
	教師	計算機の理解度や具体的な解法を確認したいとき
同意表示: AGRPRE (AGReement PREsent)	教師	計算機からの確認要求に対して肯定するとき
	計算機	教師からの確認要求に対して肯定するとき
不同意表示: DISPRE (DISagreement PREsent)	教師	計算機からの確認要求に対して否定するとき
	計算機	教師からの確認要求に対して否定するとき 問題文からコマンド候補が全く抽出できないとき
情報要求: INFREQ (INFormation REQuest)	計算機	コマンドの具体的なパラメータ値が決定できないとき 計算機からの確認要求に対する教師の応答が不同意の時
	教師	計算機の理解度や具体的な解法を知りたいとき
情報提供: INFPRE (INFormation PREsent)	教師	計算から情報要求があったとき (対象物の指示や解法の教示など)
	計算機	教師から情報要求があったとき コマンドや推論手続きの実行に失敗し、その事実を伝えたいとき
言い直し要求: CORREQ (CORrection REQuest)	計算機	教師からの発話文の解析に失敗したとき 教師からの発話文の発話意図がつかめないとき
	教師	計算機から言い直し要求があったとき

3.2.4 システムの設計方針

前節までの考察に基づき、対話支援型問題解決システムを開発する上での指針を以下のように定めた。

- (1) 将来の拡張可能性を考慮して、システムをテキスト認識部、問題解決部、対話文処理部に分ける。
- (2) 問題解決部で使用されるコマンドには宣言型知識と手続き型知識が含まれており、しかもコマンド毎に問題解決や推論方式が異なるため、柔軟な制御機構が実現できるフレーム表現を採用する。
- (3) コマンドの追加や修正に対応できるように、コマンドフレーム全体の制御機構、すなわち問題解決の制御構造はできるだけ単純なものとし、細かい制御は各フレームの付加手続きで行うものとする。
- (4) 文脈情報をシステム全体の知識ベースと考え、各サブシステムの任意のモジュールから参照、修正、追加が行えるような構成とする。

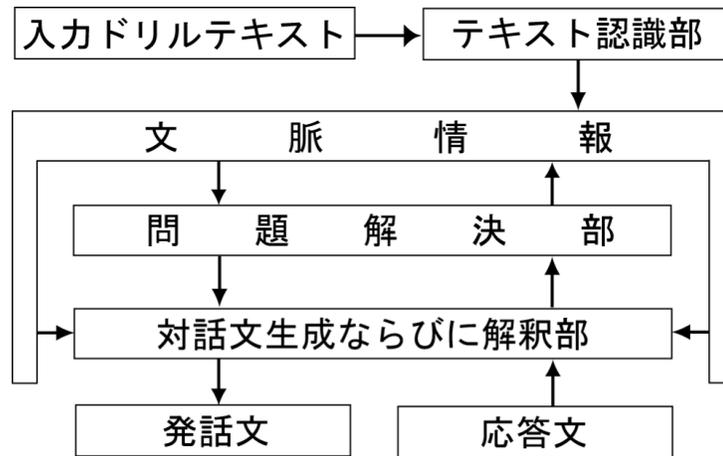


図 3.3: 対話支援型問題解決システム

(5) 知識獲得機能の追加やマルチモーダル対話への拡張が可能となるような構成とする。

3.3 システム構成と文脈表現

対話支援型問題解決システムの構成を図 3.3 に示す。テキスト認識部により、ドリルテキストの問題文は文字列から文節系列、構文木を経て格構造に、問題図形は図形モデルやテキストモデルとの照合により図形構造表現に変換される。テキスト認識部の詳細については文献 [10] と [11] を参照してほしい。問題解決部では、格構造と図形構造表現を参照して、コマンド系列を生成し、実行する。この過程で不具合が生じると、対話文生成ならびに解釈部を通して教師と対話を行い、解決を試みる。

3.3.1 言語情報の意味表現

テキスト中の言語表現（問題文）ならびに対話文の意味は次のような格構造で記述する。

```

[述語, [pred, 意味素性], [mood, 法情報],
  {[aspect, 相情報], [negative, 否定情報]},
  [格要素 1, [格ラベル 1, 意味素性],
  {[修飾要素, [(modn, moda), 意味素性],
  {[格要素, 修飾要素], ...}}}],
  [格要素 2, ...], ...].
  
```

これは 260 個の問題文および想定される対話のシナリオの分析結果に基づいて定義したものである。格構造中の “ { ” と “ } ” で囲まれた部分は任意要素である。格レベルには、主体 (subj) , 対象 (obj) , 比較の対象 (objc) , 手段 (ins) , 場所 (loc) , 場所起点 (locf) , 状態終点 (statet) , 程度 (degree) , 様態 (manner) , 形式 (form) などがある [12]。modn (moda) は連体 (連用) 修飾格を表す。また、法情報は平叙文 (declarative) , 命令文 (imperative) , 疑問文 (question) のいずれであるかを、相情報は継続相か否かを表す。図 3.4(a) に図 3.1(b) の問題文の格構造を示す。

3.3.2 図形情報の意味表現

ドリルテキストの問題図形の意味は図形構造表現と呼ぶ次のような形式で記述する。

p_form(INST, CLASS, TXT, L_TYP, LOC, STRUCT, SUB).

INST : 図形要素のインスタンス名

CLASS : 図形要素のクラス名

TXT : 図形をテキスト要素とみたときの名称

L_TYP : 図形の位置情報の指定方法

center : 図形の重心座標 $[X_i, Y_i]$

corner : 図形の頂点座標 $[X_i, Y_i, X_j, Y_j]$

LOC : 図形の位置情報

STRUCT : 部分図形表現間の空間的位置関係

(a) [ver, hor, rnd, inside, i]

(b) [inside, [..., [row(i), cn_i, [..., [column(j), k], ...], ...]]

SUB : 部分図形の p_form のリストのリスト

[[..., p-form, ...], [..., p-form, ...], ...]

図 3.1 (b) の図形部分の構造表現の例を図 3.4 (b) に示す。STRUCT は SUB と連携して解釈されるもので、(a) は SUB の i 番目のリストに列挙された要素が縦方向 (ver) あるいは横方向 (hor) に並ぶか、散在する (rnd) ことによって図形 INST を構成するか、または INST が SUB の i 番目のリストの要素を包含していることを示している。一方、(b) は図形が表を含む場合の表現法であり、表の i 行 j 列の欄の内容が SUB の k 番目のリストの要素であることを示している。cn_i は i 行目の列数を表す。

[むすぶ, [pred, link], [mood, imperative],
 [かあど, [obj, card],
 [かず, [modn, numeral], [おなじ, [moda, same]]]],
 [せん, [ins, line]]].

(a) 問題文の格構造

p_form(図形 (1), figure, 図形表現, [], [], [ver, 1], [[
 p_form(図形 (2), figure, 上連結図形群, [], [], [hor, 1], [[
 p_form(図形 (3), figure, 上連結図形, [], [], [ver, 1], [[
 p_form(円 (1), circle, 部分問題番号, center, [x1, y1],
 [inside, 1], [[p_form(文字 (1), char, char, あ, [])]]),
 p_form(矩形 (1), rectangle, 問題図形, corner, [x3, y3,
 x4, y4], [inside, 1], [[p_form(ちょう (1),
 butterfly, picture, center, [x5, y5], [], [])]]),
 p_form(中黒点 (1), dot, 連結点, center, [x6, y6], [], [])]],
 p_form(図形 (4), ...), p_form(図形 (5), ...), ...]],
 p_form(図形 (7), figure, 下連結図形群, [], [], [hor, 1], [[
 p_form(図形 (8), figure, 下連結図形, [], [], [ver, 1], [[
 p_form(中黒点 (5), dot, 連結点, center, [x7,y7], [], []),
 p_form(矩形 (5), rectangle, かあど, corner, [x8, y8,
 x9, y9], [inside, 1], [[p_form(数 (1), numeral,
 numeral, center, [x10, y10], 0, [])]]]])),
 p_form(図形 (9), ...), p_form(図形 (10), ...), ...]]]]).

(b) 図形構造表現

図 3.4: ドリルテキストの意味表現

3.3.3 コマンドの表現

表 3.1に示すようなコマンドは，システム内では次のようなフレーム形式で表現する．
command_frame(コマンド名，スロット名，ファセット，値)．

コマンド名が同一の command_frame 形式の集合を，以下ではコマンドフレームと呼ぶことにする．スロットの種類は以下の通りである．

- (1) 上位フレーム (is_a) : コマンド間の階層関係．
- (2) 印字名 (pname) : システムからの発話文生成用．
- (3) パラメータ数 (parameter_no) ．
- (4) 優先順位 (priority) : 組立・実行の優先順位
- (5) 格枠 (case_frame) : 発話文生成用
[文型 n,v,[述語印字名,[格ラベル, パラメータ No, 格助詞],
[格ラベル, パラメータ No, ...], ...].
- (6) 共起コマンド (collocation) : このコマンドと常に対の形で用いられるコマンド．
- (7) コマンドの出所情報 (source) : text (問題文の格構造) , dialogue (教師からの発話文の格構造) , inference (付加手続きによる推論) など．
- (8) コマンドの状況 (status) : incomplete (未完成) , complete (完成) , executed (実行済) など．
- (9) 手続き名 : 問題文や教師からの応答文の格構造よりパラメータを抽出する手続き (com_para) , コマンドの組立・実行のためにコマンド毎に定義された付加手続き (com_para) など．
- (10) コマンドパラメータ (parameter(i)) : このスロットは以下のファセットをもっている．
 - (a) タイプ (type) : 図形表現 , 概念 , 数値 , 手続き名 , 実行可能性などパラメータの型を表す．
 - (b) 概念 (concept) : 言語表現と図形表現の対応付けに用いるもので , 意味素性で指定する．
 - (c) 印字名 (pname) : システムからの発話文生成用．

```

command_frame(link, pname, value, むすぶ).
command_frame(link, parameter_no, value, 2).
command_frame(link, priority, value, 3).
command_frame(link, case_frame, value, [v, [むすぶ, link],
                                         [objc, 2, と], [obj, 1, を], [ins, せん, で]]).
command_frame(link, com_para, procedure, link_para).
command_frame(link, com_proc, procedure, link_proc).
command_frame(link, parameter(1), type, p_form).
command_frame(link, parameter(1), constraint,[l_typ, center]).

```

図 3.5: クラスコマンドフレームの例

- (d) パラメータ値 (value) : 具体的な値 .
- (e) 役割 (role) : 格フレームの格ラベルとの対応付けを示すもので, 文生成に利用する .
- (f) パラメータの制約条件 (constraint) .
- (g) リンク (link) : 本パラメータと他のコマンドのパラメータとの共用関係を示す .

コマンドフレームはクラスとインスタンスに分けられる . クラスコマンドフレームはコマンドに関する知識を記述したもので, 事前に定義しておく . 図 3.5 に link コマンドのクラスフレームの例を示す . 一方, インスタンスコマンドフレームはクラスフレームに基づいて問題解決過程で生成されるもので, コマンド名には link(1) のように, インディケータを付加して区別する .

3.3.4 文脈の表現

文脈を離散的な状態 S_i で表現する . 状態は発話や問題解決過程の進展に応じて変化していくが, ここでは S_i を文 U_{i+1} が発話される直前の状態であると考え . すなわち, 状態 S_{i-1} で U_i を発話すると, 状態は S_i に変化する . 状態変化に着目すると, 文脈 S_i の構成要素は大きく次の 2 種類の情報に分類できる .

$$S_i = \{V_i, C\}$$

V_i は対話や問題解決の展開に応じて変化する構成要素を, C は変化しない要素を表している .

$$V_i = \{UT_i, F_i, CM_i\}$$

UT_i は状態 S_i の時点までに発話された, 先行対話文に関する情報 UI_i のすべてのリストであり, UI_i は状態 S_{i-1} で発話された文に関する情報である.

$$UT_i = \{UI_0, \dots, UI_{i-1}, UI_i\}$$

$$UI_i = \text{utterance}(i, SP_i, HE_i, U_i, CS_i, I_i, PR_i, INT_i)$$

i : 発話番号

SP_i : 話し手 (s: システム, t: 教師)

HE_i : 聞き手 (s: システム, t: 教師)

U_i : 発話文 (表層構造, 単語のリスト)

CS_i : 発話文の格構造

I_i : 発話意図 (表 3.2を参照のこと)

PR_i : [RU, REL] 発話 U_i と対になる発話

RU: 対になる発話番号

REL: 以下に示す関係名

情報要求 情報提供 (INFREQ-PRE)

確認要求 同意表示 (CONREQ-AGR)

確認要求 不同意表示 (CONREQ-DIS)

言い直し要求 言い直し (CORREQ-PRE)

INT_i : [[L, C, D], ...] 発話 U_i 中の事物表現と現場の事物 (ドリルテキストの図形構造表現の要素など) との対応付けを示すリスト.

L: 言語表現 (単語のリスト), C: 概念表現 (格構造),

D: 事物 (図形構造表現など)

F_i は文脈 S_i の時点までの話し手 (システム) の注意状態 FE_i の変化リストである.

$$F_i = \{FE_0, FE_1, \dots, FE_{i-1}, FE_i\}$$

$$FE_i = \text{attention}(i, C_COM_i, F_ELM_i)$$

i : 注意番号

C_COM_i : 現在組立・実行中のコマンド名

F_ELM_i : 焦点要素

CM_i は文脈 S_i の時点までに問題文の格構造または教師の発話文から抽出された, あるいは推論されたインスタンスコマンドフレーム群とこれらに対応するクラスコマンドフレーム群である. この情報は対話や問題解決の進展に応じて頻りに修正や追加が行われる. 一方, 変化しない文脈の構成要素 C には, CS (問題文の格構造), PF (問題の図形構造

表現) , DIC (単語辞書) , CPT (概念の階層構造) などがある .

3.4 問題解決過程と文脈処理

ドリルテキストの問題解決過程は (1) 文脈情報の初期値 S_0 の設定 , (2) 問題解決用コマンドの組立・実行のプランニング , (3) 選択されたコマンド (カレントコマンド) の付加手続きの実行からなる .

3.4.1 文脈情報の初期値設定

初期化の主要な作業は , 当該テキストの問題解決に必要なとなるインスタンスコマンドフレーム群の生成である . まず , 問題文の格構造 (CS) 中の次のような情報を手がかりにしてコマンド候補を抽出する .

(a) 格構造の述語 : write write など

(b) 格構造の述語と特定の格要素 :

become, statet=how-many count

paint, obj=color paint

enclose, ins=circle enclose

link, ins=line link など

(c) 格要素 : number(modn numeral) count

(d) 修飾要素 : moda=same, mobj=numeral eqn

(e) 格ラベル : degree repeat など

コマンド候補が決定されると , 対応するクラスコマンドフレームの知識を利用してインスタンスコマンドフレームを生成する . さらに , コマンドパラメータ抽出手続き com_para を用いて決定可能なパラメータをすべて取り出す . 図 3.4 (a) の格構造から生成されたコマンドフレームの一部を図 3.6 (a) に示す .

3.4.2 プランニングとコマンド付加手続き

格構造から抽出されたコマンド候補群の優先順位と組立・実行状況に基づいて , 次に処理すべきコマンド名を決定し , その付加手続きに制御を渡す . 以後は付加手続きのプログラムが実行される . プログラムが終了すると制御は再びプランニングの手続きに戻され , 同

様の処理を繰り返すことになる。付加手続きはコマンド毎に若干異なる部分もあるが、基本的には次のような流れで実行が進む。

- (1) コマンド候補の出所情報による確認： source スロットの値が dialogue か inference の場合は、コマンド生成が妥当かどうかを確認するために対話処理手続き（対話文生成ならびに解釈部）を呼び出す。
- (2) パラメータスロットの値の決定。
 - (2-1) 値ファセットが空で、しかも推論可能な場合には推論手続きにより値を求め、この処理が妥当かどうかを確認するために対話処理手続きを呼び出す。
 - (2-2) 値は空であるが、型が実行可能となっている場合は、コマンドの実行により値を決定する。他のコマンドの共用パラメータにも同じ値をセットする。
 - (2-3) 無条件で値が空ならば、対話処理部手続きを呼び出し、情報要求の発話を行う。
 - (2-4) 値はセットされているが、制約条件を満たさない場合は、対話処理手続きを呼び出し、情報要求する。
- (3) 共起コマンドがあれば、そのインスタンスフレームを生成し、パラメータ抽出手続きを実行する。
- (4) 図形操作を含むコマンドを実行する。

上記の手順における対話処理手続きは、(1) 発話用パラメータ（発話意図、要求事項、発話要求コマンド名）のセット、(2) 文生成ルーチンの起動、(3) 応答文入力・解釈ルーチンの起動、(4) コマンド固有の応答文の解釈処理という流れで行われる。

3.4.3 発話文の生成

文脈情報を考慮した文生成の流れを以下に示す。

- (1) 発話文の格構造の生成。
 - (1-1) 当該コマンドの case_frame スロットから格枠を、パラメータスロットから格要素を取り出す。
 - (1-2) 要求事項および役割情報を基に、未知の格要素を適当な疑問詞に置き換える。
 - (1-3) 発話意図を基に、法情報を決定する。
- (2) 格構造から表層文を生成。

(2-1) 法情報および文型から語尾情報（「（の）ですか」，「（の）ですね」など）を決定する．

(2-2) 省略・指示・提題などの修辞処理を施す．

- ・ 焦点要素 F_ELM_{i-1} とある格要素が一致し，
- ・ 述語が直前の発話と一致するならば，この格要素を省略する．（ただし，直前の発話が「同意／不同意表示」などのように，新しい事物や事象を導入する表現でない場合は，その前の発話を直前の発話と見なす．以下同様）
- ・ 述語が直前の発話と一致せず，しかも発話意図が「確認要求」ならば，この格要素を代名詞「これ」や「ここ」などで置き換える．（ただし，焦点要素が発話ではなく問題解決（コマンド実行など）で得られたものならば，この格要素の直前に連体詞「この」を付加する．）
- ・ 述語が直前の発話と一致しないならば，この格要素を提題要素とし，文頭に移した後，格助詞の直後に係助詞「は」を付加する．（ただし，格助詞が「が」と「を」の場合は，助詞「は」で置き換える．）
- ・ ある格要素が F_ELM_{i-2} 以前の焦点要素と一致するならば，この要素の直前に適当な修飾語（「さきほどの」，「さっきの」など）を付加する．

(2-3) 文型に応じて語順を決定する．

(3) 発話を実行（ U_i を画面に出力）．

(4) 文脈情報 UI_i ， FE_i のセット．

3.4.4 応答文の解釈

文脈情報を考慮した文解釈の流れを以下に示す．

(1) 応答文を読み込む（文脈情報 U_i のセット）．

(2) 応答文を構文・意味解析して格構造（単一文種での解析の結果得られるもので，述語や一部の格要素が未決定であったり，曖昧性などが残っている構造）を求める．

(3) 格構造の意味解釈（狭義の文脈処理）．

(3-1) 応答文の発話意図の推定：例えば，発話意図が「情報要求」で，しかも対になる発話が未決定の発話が文脈情報の中にあり，かつ応答文の文型が（並列）名詞型か格要素型，法情報が「平叙文」の場合，本発話の意図を「情報提供」とみなす（文脈情報 I_i と PR_i のセット）．

- (3-2) 省略要素の推定：一般に，情報提供は要求の発端となったコマンドに対応する格構造中の要求事項に対してなされる．そこで，名詞型（格要素型）表現を要求格要素に代入することで，格構造を復元する（文脈情報 CS_i のセット）．
- (3-3) 応答文中の事物表現（表層構造，格構造）に対応する図形構造表現（PF）中の要素を同定する（文脈情報 INT_i のセット）．
- (3-4) 直前の発話にない要素が現れたときは焦点を更新する（文脈情報 FE_i のセット）．
- (3-5) 応答文の格構造から新たなコマンド候補が抽出できるかどうかチェックする．抽出できる場合は，インスタンスコマンドフレームを生成し，パラメータ抽出手続きを実行する．

3.5 実験と考察

対話支援型問題解決の実験システムをワークステーション上の Prolog を用いて実現した．

3.5.1 対話例

図 3.1 (b) のテキストの問題解決に対する対話と文脈処理を考えてみよう．問題文の格構造（CS，図 3.4 (a)）と図形構造表現（PF，図 3.4 (b)）が与えられて，問題解決過程が始まる．図 3.6 (a) が CS から抽出されたインスタンスコマンドフレーム群 CM_0 である．コマンドの優先順位と状況により， $eqn(1)$ がカレントコマンドとなり，その付加手続き eqn_proc が起動される．この時点における， $C = \{CS, PF, DIC, CPT\}$ ， $V_0 = \{UT_0(= \quad)\}$ ， $F_0 = \{attention(0, eqn(1), [])\}$ ， CM_0 から構成される S_0 が文脈情報の初期状態である．

手続き eqn_proc はコマンド $eqn(1)$ の第 2 パラメータの概念および値ファセットが未知であることから，発話意図「情報要求」を生起させ，図 3.7 の発話文 S_1 を生成する．発話文 S_1 の表層構想 U_1 ，格構造 CS_1 ，発話意図，カレントコマンド，話し手と聞き手の情報などを付加した状態を S_1 とする．

システムからの情報要求に対して，教師から応答された文が T_1 である．この文の解釈を文脈 S_1 の下で試みる．文型は名詞型表現であるが， CS_1 を参照して省略された述語が「おなじ」と推定し，格構造 CS_2 を生成する．また， S_1 の発話対の有無，応答文の文型，法情報から T_1 の発話意図を S_1 に対する「情報提供」とする．次に，「ひだり

うえの...ちょう」という言語表現に対応する個体値を文脈 S_1 の PF 中から同定し、これを新たな焦点要素とする。こうして得られた文脈 S_2 の要素 UI_2 と FE_2 を図 3.8 に示す。

文 T1 の解釈の過程で格構造からの新たなコマンド抽出の可能性がチェックされる。objc 要素の意味素性が “number” であることから、コマンド候補 count が抽出され、図 3.6(b) に示すインスタンスコマンドフレーム count(1) が生成される。Count(1) の第 2 パラメータの「かず」は eqn(1) コマンドの第 2 パラメータにも対応している。この関係を示すためのスロットが図 3.6 (b) の最後の 2 行である。付加手続き実行中に、新たなコマンドが生成された場合は、トップレベルに戻り再度プランニングをやり直す。この結果 count(1) がカレントコマンドとなり、付加手続き count_proc に制御が移る。

手続き count_proc がコマンド count(1) の抽出の妥当性を確認するために発話した文が図 3.7 の S2 である。ここで、焦点要素が「...ちょうのかず」で、objc 要素と一致し、発話意図が「確認要求」であるため、この要素を代名詞「これ」で置き換えている。S2 に対する応答が T2 であり、発話意図は「同意表示」である。これによりコマンド count(1) が実行可能となり、これを実行後、第 2 パラメータを決定する。この結果は、パラメータの共用リンクをたどって eqn(1) の第 2 パラメータにもセットされる。その後、制御は再び eqn_proc に戻される。

手続き eqn_proc では、eqn コマンドを実行し、その第 1 パラメータに値をセットする。次に、eqn がオペレーションではなく述語であるため、共起のオペレーションコマンドである select が推論され、インスタンスフレームが生成される。この様子を図 3.6 (c) に示す。eqn_proc は処理を終了し、プランニングを経て、カレントコマンドは select(1) になる。

付加手続き select_proc が select(1) の推論の妥当性を確認するために発話した文が S3 である。select(1) コマンドに eqn(1) 述語が埋め込まれているため、フレームから 2 つの格枠が取り出され、発話文も埋め込み文となる。述語「おなじ」の objc 要素は、文 S2 と同じ理由で代名詞に置き換えられている。次に、select_proc は第 1 パラメータの概念と値ファセットが未知であることから、発話意図「情報要求」を生起させ、発話文 S4 を生成する。これに対する応答文 T4 により、第 1 パラメータの値が定まり、select(1) が実行可能となる。コマンド実行の後、第 3 パラメータが定まり、その値 p_form(矩形(6), ...) が新たな焦点要素となる。コマンドフレームの変化を図 3.6 (d) に示す。

最後に残った link(1) がカレントコマンドになり、制御は link_proc に移る。この発話文が S5 と S6 であり、図 3.6 (e) が最終的なコマンドの状態である。

`command_frame(eqn(1), source, value, text).`
`command_frame(eqn(1), status, value, incomplete).`
`command_frame(eqn(1), com_proc, procedure, eqn_proc).`
`command_frame(eqn(1), parameter(1), concept, numeral).`
`command_frame(eqn(1), parameter(1), pname, かず).`
`command_frame(eqn(1), parameter(1), value, []).`
`command_frame(eqn(1), parameter(2), concept, []).`
`command_frame(eqn(1), parameter(2), value, []).`
`command_frame(link(1), parameter(1), pname, かあと).`
`command_frame(link(1), parameter(1), value, []).`
`command_frame(link(1), parameter(2), concept, []).`
`command_frame(link(1), parameter(2), value, []).`

(a) コマンドフレームの初期状態 CM_0

`command_frame(count(1), com_proc, procedure, count_proc).`
`command_frame(count(1), source, value, dialogue).`
`command_frame(count(1), parameter(1), concept, butterfly).`
`command_frame(count(1), parameter(1), value,`
`p_form(矩形(1), rectangle, 問題図形, ...)).`
`command_frame(count(1), parameter(2), type, executable).`
`command_frame(count(1), parameter(2), link, [eqn(1), 2]).`
`command_frame(eqn(1), parameter(2), link, [count(1), 2]).`

(b) コマンドフレームの状態 CM_2

`command_frame(count(1), parameter(2), value, 1).`
`command_frame(count(1), status, value, executed).`
`command_frame(eqn(1), parameter(2), value, 1).`
`command_frame(eqn(1), parameter(1), value, 1).`
`command_frame(select(1), source, value, inference).`
`command_frame(select(1), com_proc, procedure, select_proc).`
`command_frame(select(1), parameter(1), value, []).`
`command_frame(select(1), parameter(2), pred, [eqn(1)]).`
`command_frame(select(1), parameter(3), pname, かあと).`
`command_frame(select(1), parameter(3), value, []).`
`command_frame(select(1), parameter(3), link, [link(1), 1]).`

(c) コマンドフレームの状態 CM_4

```

command_frame(select(1), parameter(1), concept, figure).
command_frame(select(1), parameter(1), value,
               p_form(図形 (7), figure, 下連結図形群, ...)).
command_frame(select(1), parameter(3), value,
               p_form(矩形 (6), rectangle, かあど, ...)).
command_frame(select(1), status, value, executed).
command_frame(link(1), parameter(1), value,
               p_form(矩形 (6), rectangle, かあど, ...)).
command_frame(link(1), parameter(2), concept, butterfly).
command_frame(link(1), parameter(2), value,
               p_form(矩形 (1), rectangle, 問題図形, ...)).
(d) コマンドフレームの状態 CM8    command_frame(link(1),
parameter(1), concept, dot).
command_frame(link(1), parameter(1),, value,
               p_form(中黒点 (6), dot, 連結点, ...)).
command_frame(link(1), parameter(2), value,
               p_form(中黒点 (1), dot, 連結点, ...)).
(e) コマンドフレームの状態 CM12

```

図 3.6: コマンドフレームの生成過程 (続き)

- S1: なんと おなじ かずですか。
- T1: ひだりうえの わくの なかの ちょうの かずです。
- S2: これを かぞえるのですね。
- T2: そうです。
- S3: これと おなじ かずの かあどを えらぶのですね。
- T3: そうです。
- S4: どこから えらぶのですか。
- T4: したにある 4つの ずけいからです。
- S5: さきほどの ちょうと このかあどを
せんで むすぶのですね。
- T5: そうです。
- S6: どこと どこを むすぶのですか。
- T6: ちょうの したにある くらいてんと
かあどの うえにある くらいてんです。

図 3.7: システムと教師の対話例

$UI_1 = \text{utterance}(2, t, s, U_2, CS_2, I_2, PR_2, INT_2)$
 $U_2 = [\text{ひだりうえ}, \text{の}, \text{わく}, \text{の}, \text{なか}, \text{の},$
 $\text{ちょう}, \text{の}, \text{かず}, \text{です}, \text{。}]$
 $CS_2 = [\text{おなじ}, [\text{pred}, \text{same}], [\text{かず}, [\text{subj}, \text{numeral}]],$
 $[\text{かず}, [\text{objc}, \text{number}], [\text{ちょう}, [\text{modn}, \text{butterfly}]],$
 $[\text{なか}, [\text{modn}, \text{inside}], [\text{わく}, [\text{modn}, \text{rectangle}]],$
 $[\text{ひだりうえ}, \text{modn}, \text{upper-left}]]]]]]$
 $I_2 = \text{INFPRE} \quad PR_2 = [1, \text{INFREQ-PRE}]$
 $INT_2 = [[[\text{ひだりうえ}, \text{の}, \text{わく}, \text{の}, \text{なか}, \text{の}, \text{ちょう},$
 $[\text{ちょう}, [\text{concept}, \text{butterfly}], \dots]]$
 $p_form(\text{矩形 (1), rectangle, 問題図形, } \dots)]]$
 $FE_2 = \text{attention}(2, C_COM_2, F_ELM_2)$
 $C_COM_2 = \text{eqn}(1)$
 $F_ELM_2 = [\text{ひだりうえのわくのなかのちょうのかず}]$

図 3.8: 文脈情報の例

3.5.2 対話処理部の評価

図 3.1の例を含む 7 種類のテキストに対して行った．実験データの詳細と評価結果は次のとおりである．

- (1) 被験者数：4 名（システムの開発に直接関わっていない教官 1 名と大学院生 3 名）．
- (2) 問題数：7 題
- (3) 対話数：28（各被験者が各問題ごとに 1 対話）．
- (4) 1 対話の平均発話対（ターン）数：6.4
- (5) システムからの平均発話長：文節数 3.2，文字数 15.6．
- (6) 教師からの平均発話長：文節数 3.0，文字数 12.1(教師からの発話の 51 %を占める同意表示の「はい」や「そうです」などを除くと，文節数 5.1，文字数 20.7 となる)．
- (7) 言い直し要求の相対的頻度：教師からの発話文 179 文に対して 19 文．
- (8) 言い直し要求の典型的な例：
 - (a) 未知語の存在：「『・・・』ということばがわかりません」(9 例)
 - (b) 構文・意味解析の失敗「ぶんの かいせきができません」(5 例)
 - (c) 名詞のテキスト中での指示対象の同定ミス「・・・がみつかりません」(5 例)
- (9) 被験者による主観評価の結果
 - (9-1) システムからの発話文の評価
 - (a) 了解性：問題なし(4 名)，問題あり(0 名)
 - (b) 自然性：問題なし(2 名)，問題あり(2 名．発話文が生硬すぎる，くどい．指示詞がわかりにくい)．
 - (9-2) 応答文の入力に対する評価
 - (a) 入力の容易性：容易(1 名)，容易でない(3 名．テキスト上の対象物を言語のみで表現するのが難しい．問題解決過程が対話のみでしか表現されないのので，適切な応答文を作るのが難しい．)
 - (b) 解析能力：高い(0 名)，普通(2 名)，低い(2 名．複雑な係り受けの解析ができない．漸増的な解析ができないので，1 カ所でも誤りがあると全て言い直し要求させられる)．

以下では、これらの結果と問題点について考察を行う。第 1 は、指示詞の問題である。修辞処理の指示詞については、文脈指示を想定しており、しかも「こ」系列のみを扱っている。対話の現場にはテキストが存在しており、指示の大部分はテキストの要素に対して行われるため、文脈指示に加えて現場指示も考慮する必要がある [34]。さらに「こ、そ、あ、ど」の 4 系列を使い分けも考えなければならない。現在、ポインティングジェスチャを併用した対話処理方式の開発を行っており、これを併せて話し手、聞き手との位置関係を基準とした発話文の指示処理を検討中である。

第 2 の問題は応答文の入力方法で、システムからの情報要求に対して、この応答を文章だけで表現するのが難しいということである。図表を含む算数のドリルテキストの問題解決を教示するといった状況においては、一般に指示や描画などのジェスチャを交えながら対話を行っており、これを自然言語のみで行おうとするところに入力の困難さがあるものと思われる。さらに、キーボードによる応答文入力の煩わしさも大きな問題である。この問題の解決法については、カメラから入力された一連のジェスチャ動画像と発話文との協理解システム [13] 並びに音声発話文解析処理方式 [25] の開発を進めており、本問題解決システムに統合する手法を検討中である。

第 3 の問題は応答文の解析能力である。現在のシステムで扱える応答文の種類はドリルテキストに準じたものに限られており、制限が大きいと思われる。この問題については音声発話文解析処理方式の導入と併せて改良を行う予定である。

3.5.3 試作システムの問題解決能力

3.2.1 で述べたように現在 30 個のコマンドを抽出しており、そのうち 15 個を表 3.1 に示している。ドリルテキスト [52] および [53] に記述されている 260 個の問題の中で、30 個のコマンドをいくつか組み合わせかつ教師との対話を行うことにより取り扱えるものは 217 題である。残り 43 個の問題は、複雑な図形の表現に対する計測や形状の類似性判定に関わるものであり、これらの問題解決には、より粒度の細かいコマンドの設定や図形認識能力の向上が必要になるとと思われる。また、解決可能な 217 題の中の 62 題の純粋な文章題については、その解決部、すなわち文章表現のみからコマンド系列の生成する部分が現在のシステムには組み込まれていない。

本節の始めで問題解決システムの諸条件について述べたが、これらの条件、特にコマンドパラメータの制約条件を変えるとシステムの振る舞いに違いが生じる。主な違いは、システムからの発話文に現れる。例えば、制約条件を強くするとパラメータ抽出手続きが有効に働くことになり、情報要求の発話が減少し、確認要求の発話が増加する。

3.5.4 対話支援型問題解決システムの意義

本システムでは、算数の問題解決のための要素的処理手続きをコマンドフレームで表現し、カレントコマンドの付加手続きの実行中に不都合が生じると、このフレームに埋め込まれた格枠を利用して発話文を生成し [43]、教師からの応答を求める。従って、事前に定義された知識が不十分な場合でも、対話を通して問題を解決することができる。さらに、対話処理では、文脈情報を用いて自然な発話が行えるように工夫している。本システムでは対象分野を小学校 1 年算数のドリルテキストとしたので、システムが生徒の役割、人間が教師の役割となっている。しかし、本システムの目的は、生徒の役割を果たすシステムを構築することではなく、自然言語による対話処理機能を備えた問題解決システム実現の手がかりを得ることにある。したがって、算数や理科の問題に限らず、経営戦略や株式投資、ロボットの経路制御などに関する問題が考えられる。本システムで使用しているコマンドフレームを定義し直すことにより、本枠組みは他分野の問題解決にも適用可能であると思われる。

しかしながら、対話に基づく問題解決手法が有効となるためには、対話から得られた知識の再利用、すなわち知識獲得が可能でなければならない。対話により獲得される知識には、教師のもつ問題解決のための以下のような経験則的知識が考えられる。

- (1) コマンドのパラメータの各種制約条件
- (2) コマンドの共起関係のパターン
- (3) 問題のタイプ毎のコマンド系列のパターン

このような知識獲得機能の追加については、3.2.4で述べたシステム設計方針の(2)および(4)とも関係するが、各種知識をフレーム形式で表現し、文脈情報をシステム全体の知識ベースと考えることにより可能になると考えている。具体的に示すと、上記(1)と(2)については、当該インスタンスコマンドに、対話から得られた情報を保存するスロットを設けることにより対処できる。一方、(3)のパターンを表現するためには、新たに問題解決フレームと呼ぶような上位の構造を導入する必要があるだろう。このフレームは具体的な問題毎に生成され、この種の問題解決に必要なコマンド名を値とするスロットをもつことになる。問題タイプの同定のために、問題文の格構造ならびにコマンド候補列を保持するスロットを準備しておき、一種の事例ベース推論システムとして機能するような構成が考えられる。現在、これらのフレームの詳細な定義と処理手続きの検討を進めている [12]。

3.5.5 言語表現と図形表現の統合化

本問題解決システムの特徴の一つは図形表現が言語表現と相補的に機能する形で提示される問題を取り扱えることにある。このことが発話文生成機構に与える制約としては、言語表現（問題文）には現れず、図形表現のみに現れる各種オブジェクトやこれらの相互関係を言語表現に変換して発話しなければならない点が挙げられる。この点は、逆に教師の発話文解釈において言語表現された事物を図形表現上のオブジェクトに対応させる場合にも関係してくる。現システムではテキストの図形部分の解釈結果を図形構造表現で記述している。これは、テキスト中に現れる全ての図形オブジェクトの名前、種類、座標及び相互の位置関係を記号表現したものであり、問題文や対話文の格構造と容易に対応づけることができる。したがって、現システムの発話文生成機構は、図形表現を用いない問題解決システムのそれと基本的に大きく変わることはない。

しかしながら、「 で囲む」や「数字を書く」といった解答合成操作の一部には、図形構造表現には現れない背景部分の位置の指定を必要とするものがある。さらに、自然言語による発話に加えて、ポインティングジェスチャなどを含めたマルチモーダル対話においては、図形に2次元的な空間表現が不可欠である。このためには、画像データと図形構造表現の中間に位置し、任意の図形オブジェクトの位置や空間的關係が直接記述可能な表現方式や、発話とジェスチャの意味を表現する枠組みが必要となる。このようなマルチモーダル対話への拡張についても、対話文生成・解釈部を問題解決から独立させていること、各種知識をフレーム形式で表現し、文脈情報として独立させていることで可能になると考えている。現在、画像データ中の各オブジェクトのラベル情報を中間表現として文脈情報に追加し、発話文の格構造をジェスチャ情報が記述できるように拡張することで、マルチモーダル出力の検討を進めている [26]。

3.6 まとめ

小学校1年算数のドリルテキストおよび教師との対話を相互参照しながら問題解決を行うシステムにおける文脈情報の表現法とこれに基づく対話文の生成・解釈法を提案した。対話処理実験システムは Prolog の下で稼働しており、文脈処理と発話文生成機構が良好に動作することを確認している。この結果、コマンドの組立と実行からなる問題解決過程の手続きの流れならびに文脈情報と対話の言語表現を対応付ける手がかりが得られた。本対話制御方式は、システムが行う問題解決や知識獲得過程をどのような時点で、どのような発話文を生成すればよいか、またそれに対応する応答文を、文脈情報を用いてどのように解釈すればよいかという点に重点を置いており、システム主導の対話処理機能を備えた

問題解決システム実現の一助になることが期待できる。

今後の課題としては、(1) 対話により得られた情報の再利用(知識獲得)法の検討、(2) 音声対話処理部の導入、(3) 発話文とジェスチャを併用したマルチモーダル対話の実現などが挙げられる。

第 4 章

表からの情報抽出

Web 上の文書は文章だけでなく、表や画像なども用いられている。しかし、文章と表、文章と画像といった異なった形式で表現された情報やメディア間の関係を考慮した研究はあまりなされていない。また、情報を提示するという観点から考えると、表には文章や画像では表現できない情報が、画像には文章や表では表現できない情報が含まれており、文章だけの解析では十分な検索結果や要約が得られるとは限らない。我々は複数のメディアを統合・処理するという目的から、まず表を解析する意義とその活用法について論じる。具体的には、表からの情報抽出と得られた結果を用いた検索・統合・要約処理について考察する。その有効性を示すために PC を対象とした製品選択支援システムを構築する。

4.1 はじめに

近年のインターネットの急速な普及により、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、紙面で伝えられていた情報の電子化が進んでいる。電子化されるようになった情報の 1 つとして、図 4.1 に示すような製品の性能や機能などを記述した表（以下、製品性能表と呼ぶ。）が挙げられる。製品性能表には、その製品に関する様々な具体的なデータが載せられている。しかし、データが載せられているだけで、どの項目がその製品の特徴となるかは製品性能表を見ただけでは一般的には判断しにくい。あるユーザが PC を購入しようとしていると仮定すると、そのユーザは多くの PC メーカーのサイトから製品性能表などを含んだ様々な情報を抽出し、比較する必要がある。しかしながら、多数の製品性能表から求めている製品を探し出すことは一般に労力がかかる。その理由としては、

1. 各サイトでは自社製品の特徴は述べられているが、他社製品などとの比較はあまりなされていない。

2. 各サイトごとに様々な表現方法がある。
3. 要求と製品の特徴を関連づけるには、その製品に対してある程度知識が必要である。

などが挙げられる。ユーザの要求を満足させるには、複数のサイトから情報を抽出し、統合する必要がある。また、各々の製品の相対的な特徴正しく抽出できたとしても、表を提示するだけでは、ユーザにとって読みやすい情報であるとは限らない。

我々は、複数の製品性能表を解析し、各々の製品データを比較することで特徴データを抽出し、その特徴データを用いたランキング、文章や表、グラフといった複数の形式で出力する製品選択支援システムの開発を進めている [16][57]。本章では、表形式で記述された製品性能表の抽出・解析手法、相対的な製品特徴の抽出法とランキング、よりわかりやすい形でのユーザへの情報提示手法などについて述べる。図 4.2 に本システムの概要を示す。図中の下部にある画像内容の特定処理については次章で述べる。システムは HTML で記述された製品性能表を各メーカーのサイトから抽出し、それらを解析することで表を表構造へ変換する。得られた表構造中の数値データや文字データを比較し、それぞれの製品の相対的な特徴データを抽出し、ユーザの要求に応じてスコア付けを行う。その特徴データを基に文章生成と表の再構成、グラフ生成を行ない、複数の形式を統合した要約を出力する。図 4.3 は我々が開発した製品選択支援システムである。このシステムの特徴は、大きく以下の 3 つである。

1. 5 種類のユーザからの要求と属性の重み選択によるスコアリング
2. 適合性フィードバックによるスコアリング
3. 文章や表、グラフによる出力

以下、まず 4.2 節で関連研究について述べる。4.3 節では HTML で記述された製品性能表の抽出処理について、4.4 節で表の解析処理について述べる。4.5 節では製品の特徴データ抽出処理やユーザの要求に基づくスコアリング手法、適合性フィードバックの利用などについて述べ、4.6 節では抽出された特徴データを用いた文章生成処理と表の再構成、レーダチャートの生成などの出力形式について述べる。4.7 節でシステム全体について考察し、4.8 節でまとめる。

ハードウェア仕様 - Microsoft Internet Explorer

アドレス: D:\shimada\HTML_RES\HWハードウェア仕様.htm

■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 42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、ACPI 1.0	
メモリ	標準/最大 64MB/192MB(SDRAM)	64MB/192MB(SDRAM)
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FLサイドライト付きTFTカラー液晶(※1)、1,024×768ドット・65,536色
	外部ディスプレイ(オプション)(※3)	最大1,280×1,024ドット・256色
	内部ディスプレイと同時表示(※4)	最大1,024×768ドット(※2)、走査周波数 垂直60Hz
	ビデオRAM	2.5MB
	グラフィックアクセラレータ	Trident Cyber9525DVD
入力装置	本体キーボード	90キーiDADG106キー連動、Windowsキー・アプリケーションキー付き、ひらがな印刷、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(※5)
補助記憶装置(固定式)	ハードディスク(※6)	6.4GB
	ソフトウェア占有率	1.6GB
	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
	対応フォーマット(※7)	音楽CD、CD-ROM、CD-R、CD-RW、マルチセッション(PhotoCD、CDExトラ)

図 4.1: 製品性能表の例

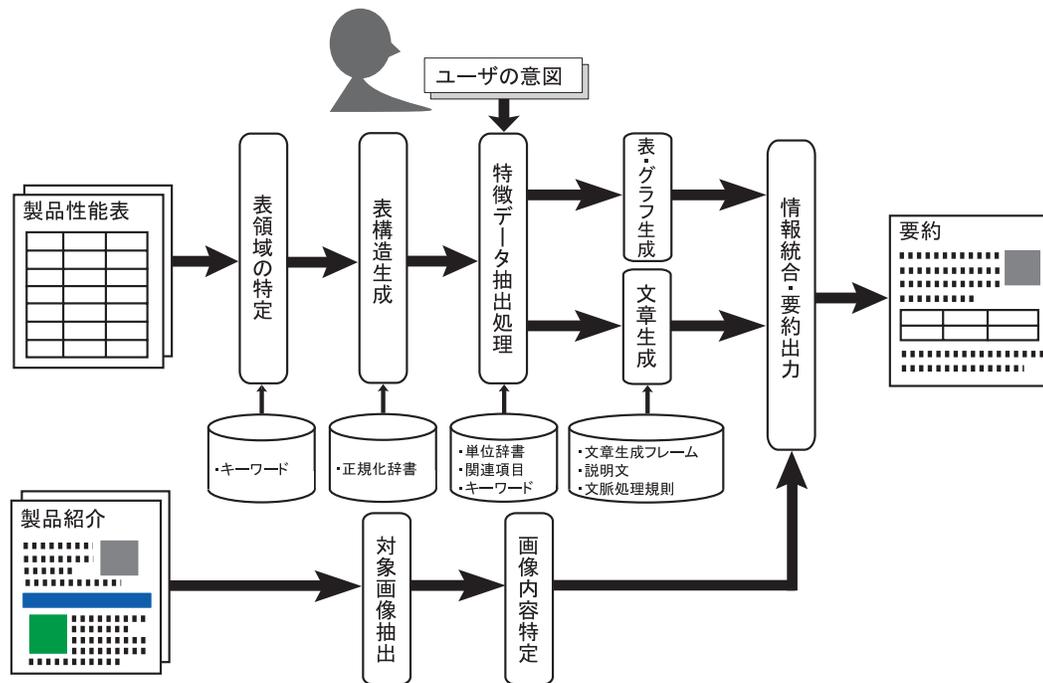


図 4.2: システムの概要

The screenshot shows a software window titled "Product Ranking System using User Requests - PRSUR- Ver.1.3.1". On the left, there is a list of filter categories: ALL, High Performance, High Graphics Performance, Mobile PC, Practical Use, and Budget. On the right, there are buttons for "Select Items", "Retrieval", "Feedback", "Summarization", "Radar Chart", "Clear", and "Exit". A "Set a ceiling on the price" dialog box is open, with a text input field and a "Yes" button. The main area is a table with columns for Rank, Model Name, Score, and Price. A context menu is open over the second row, "DynaBook DB70P/5MC", with options for "Detailed Product Information", "Selection for Radar Chart / Summarization", "Relevant Product", and "Non-Relevant Product".

Rank	Model Name	Score	Price
1	LaVie C LC800J/54ER	5.65762498503737	330000 yen
2	DynaBook DB70P/5MC	5.59770084552738	349800 yen
3	Mebius PC-RJ950R		en
4	FMV-BIBLO NE5/60C		en
5	Mebius PC-MJ700M		en
6	VAIO PCG-F76/BP		en
7	LaVie C LC60H/54DR		en
8	FMV-BIBLO NE5/800...		en
9	人 CF-X1D	4.97090473807811	249800 yen
10	Let's note CF-B5ER	4.86825449029368	279800 yen
11	DynaBook DB60C/4RA	4.86022832114343	239800 yen
12	LaVie S LS600J/55DV	4.79861152624852	299800 yen
13	VAIO PCG-XR1F/BP	4.64586396287969	249800 yen
14	ThinkPad i Series 1200	4.62003783260485	189800 yen
15	DynaBook DB55C/4CA	4.6006943757195	199800 yen
16	LaVie S LS55H/54DV	4.58063601837857	249800 yen
17	VAIO PCG-XR7F/K	4.53106173957371	279800 yen
18	VAIO PCG-F70A/BP	4.47804170419491	199800 yen
19	FMV-BIBLO MF5/55D	4.47327764683991	239800 yen
20	LaVie U LU50L/53DC	4.36736095128514	178000 yen

図 4.3: 製品選択支援システム

4.2 関連研究

製品情報などの抽出に関する研究は、単一の製品紹介の記事 [36] [33] や新聞記事 [14][55] [71][77] などの文章が対象とされることが多い。文章からの製品情報抽出は、発売日などの時間表現や販売価格などの数値表現の抽出には向いているが、各々の製品の相対的な特徴を抽出するには、情報が集約されている記事ではなく、より多くの製品データを含む一次データである表を用いる方がよい。また、文章からの抽出処理は、一般にテンプレートによるパターンマッチングで行われるが、テンプレートの作成には多大な労力がかかる。文章の書き手の違いを吸収するためにより多くのパターンを作らなければならないという問題点がある。複数製品の記事を対象とした研究もあるが [1][67]、この場合のテンプレートの数はさらに膨大になる。提案する手法は、処理対象が製品性能表であり、文章を対象とする場合よりも小規模の辞書で実現できる。同様に、文章を対象とした場合、別の製品の記事に対象を変えると、再び膨大なテンプレートを作成しなければならない。提案手法では、単位などの表層パターンをキーワードとして扱うので、文章からの情報抽出システムに比べ、処理対象を他種類のデータに移行する際に必要なテンプレート作成などの作業がスムーズに行える。

レイアウト情報を持つ文書からの内容抽出に関しては、会告記事に関する電子ニュースの自動要約を行う研究 [54] や罫線を含まない表形式の文書を対象とした研究 [28] などがあるが、その対象は箇条書きされた文章であり、ここで対象としている表とは異なる。処理対象が単一の文書であり、他の文書との比較や統合を行っていない。従来の表解析に関する研究は、文書イメージ [21] やプレインテキスト [41] を扱ったものが多い。Wang ら [73] の研究では、HTML で記述された表を扱っているが、その目的は<TABLE> タグで囲まれた領域が、本当に表であるかどうかを判定することである。HTML で記述された表の内容抽出に関する研究 [5] もあるが、処理対象は単一の表のみである。Hammer ら [18] は、HTML で記述された文書をデータベース化する手法について提案しているが、表形式については十分議論されていない。複数の表を扱う研究 [82] もあるが、その目的は表からオントロジーを抽出することである。提案する手法は、複数の製品性能表から各々の製品のデータを抽出し、そのデータを比較することで、相対的な特徴データを抽出することができる。

要約処理については、対象が文書である場合が殆どで [44][45]、文書以外のデータが用いられることは少ない。佐藤ら [54] は、箇条書きされた会告記事からダイジェストを作成しているが、出力は箇条書きのままである。梶井ら [33] は、記事から抽出した情報を表として一覧提示している。株価データや気象データを文章として要約するシステムも幾つか報告されている [29][49]。しかし、複数の製品性能表を要約する場合、読む側の視点を考えれば、出力は表だけでなく、それらを文章などと統合した出力にすることが望ましい。

WWW 上でのオンラインショッピングを手助けするシステムの報告も数多くなされている [2][3][8][9]。Doorendos ら [9] の研究は対象が価格のみで製品特徴の比較を行っていない。Budzikowska ら [2] および Chai ら [3] はユーザの意図を考慮したオンラインショッピングのシステムを提示しているが、複数社の製品情報の比較は行われていない。

本システムには次のような利点がある。

- 文章などの記事を対象とした情報抽出システムより少ないパターンや知識でシステムを実現できる。
- 情報が集約された記事ではなく、製品に関するあらゆるデータが載せられた一次データである表を扱うことで、記事からは抽出できない特徴データも抽出できる。
- 記事からの情報抽出と異なり、任意の製品群から相対的な特徴データが抽出できる。
- 特徴データ抽出処理が小規模の辞書で実現できるため、システムを他種類のデータ

に移行する際のコストが小さい。

- ユーザの要求に合致した製品を複数の表から抽出できる。
- 抽出された特徴データからの文章生成やグラフの生成，表との統合などによって，よりわかりやすい形式で出力できる。

4.3 製品性能表の抽出

まず，企業のサイトから性能表を抽出する必要がある。これらのサイトからの Web ページ抽出は既存のダウンロードソフトを用いている。本節では，まずキーワード候補の抽出処理について述べる。続いて，得られたキーワード候補に対する重み付け処理について述べる。重み付け手法としては，エントロピーを用いる手法とベイズの定理を用いる手法の二種類がある。

4.3.1 キーワード抽出

<TABLE> で記述された領域が表であるかどうかを判別するために，我々は表判別のためのキーワードを作成した。キーワードの定義は

1. 性能表の項目欄中出现する単語 (<TD> ~ </TD>)
2. 一定長以内の文章中出现する単語
3. 性能表が存在するページ内で顕著または限定的に出現する単語

とする。この定義に基づき，HTML 文書から性能表抽出のためのキーワード候補を抽出する。実際の抽出処理は以下の手順で行われる。

1. HTML 文書から<TABLE> タグで記述された領域を抽出する。
2. <TABLE> タグ中の各<TR> タグ中の初めの<TD> タグの内容を抽出する。
3. 得られた文字列が 25 文字以内であれば，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「茶釜」[35]を用いた。

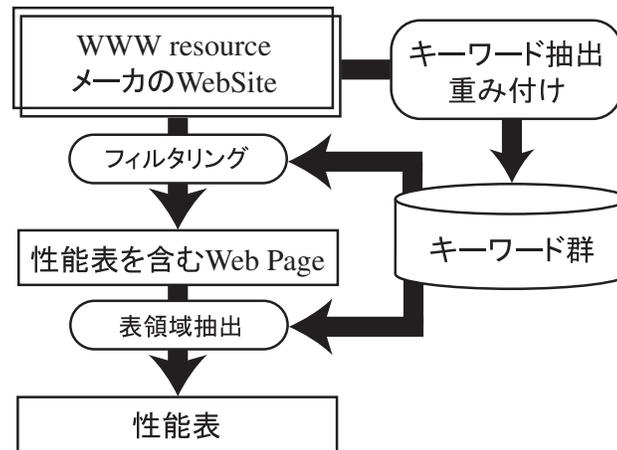


図 4.4: 性能表抽出の流れ

4.3.2 エントロピーによる重み付け

得られたキーワード候補に対して重みを計算し、その重みを基にキーワードを抽出する。情報理論的な観点から語の特定性を考えれば、ある単語が文書集合中の各文書にどれだけ偏って出現するかを情報理論のエントロピーの考え方をを用いて数値化することができる。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。キーワード抽出の手順を以下に示す。

1. HTML 文書群 $D = \{d_1, \dots, d_N\}$ からキーワード候補を抽出する。
2. キーワード候補 t の各文書 d における頻度 $tf(t, d)$ を計算する。
3. 文書群 D を性能表が存在する文書群 D_{real} とそうでない文書群 D_{no} に分割する。
4. 文書群 D_{real} におけるキーワード候補 t の重み wr_t^{real} と文書群 D_{no} におけるキーワード候補 t の重み wr_t^{no} を以下の式で計算する。

$$wr_t^{real} = \frac{w_t^{D_{real}}}{w_t^{D_{no}}},$$

$$wr_t^{no} = \frac{w_t^{D_{no}}}{w_t^{D_{real}}}.$$

ここで、

$$w_t^{D_n} = \log \sum_{k=1}^M tf(t, k) + \sum_{i=1}^M \frac{tf(t, i)}{\sum_{j=1}^M tf(t, j)} \log \frac{tf(t, i)}{\sum_{j=1}^M tf(t, j)}$$

であり、 M は文書群 D_{real} もしくは D_{no} における文書の総数である。

ここで、性能表を含む文書中の<TABLE> タグ内に顕著に存在する語をキーワードと呼ぶ。キーワードの重みは、 $ws_t^{real} = df(t, D_{real}) \times wr_t^{real}$ で求める。一方、性能表を含まない文書中の<TABLE> タグ内に顕著に存在する語をノイズワードと呼ぶ。ノイズワードの重みは、 $ws_t^{no} = df(t, D_{no}) \times wr_t^{no}$ で求める。ここで、 $df(t, D_{real})$ および $df(t, D_{no})$ は、文書群 D_{real} もしくは D_{no} 中の単語 t を含む文書の数である。

4.3.3 ベイズの定理による重み付け

ベイズの定理はパターン認識・分類の世界でよく知られた確率ベースの手法である [6] [47] [48]。 $C = [C_j]_{j=1}^M$ において、 $P(C_j)$ ($\sum_{j=1}^M P(C_j) = 1$) は事前確率と呼ばれる。事前確率と条件付き確率密度分布 $p(x|C_j)$ ($\int p(x|C_j)dx = 1$) が事前に得られる場合、単語 x がクラス C_j に属する事後確率 $P(C_j|x)$ は次の式で求められる。

$$P(C_j|x) = \frac{P(C_j)p(x|C_j)}{p(x)}$$

ここで、

$$p(x) = \sum_{j=1}^M P(C_j)p(x|C_j), \quad \int p(x)dx = 1$$

であり、

$$\sum_{j=1}^M P(C_j|x) = 1$$

となる。ここで、 $C = \{D_{real}, D_{no}\}$ である。すべての単語に対して、各クラスでの事後確率を求める。我々は上記の式で、ある単語 t を考えたとき、得られる事後確率 $P(C_j|t)$ をその単語の重みとする。すなわち、ある単語 t のキーワードとしての重みは、 $ws_t^{real} = P(D_{real}|t)$ であり、ノイズワードとしての重みは、 $ws_t^{no} = P(D_{no}|t)$ となる。

4.3.4 性能表抽出処理

図 4.4で示したように、抽出処理は、フィルタリングと表領域抽出の二つのプロセスからなる。

4.3.4.1 フィルタリング

まず、初めに HTML 文書のフィルタリングを行う。ここで、フィルタリングとは、全 HTML 文書から性能表を含む HTML 文書のみを抽出することを指す。フィルタリングでは、前節で獲得されたキーワードとノイズワードを用いて、性能表を含む HTML 文書を抽出する。フィルタリングの流れを以下に示す。

1. HTML 文書 d_i から<TABLE> タグで記述された領域を抽出する .
2. <TABLE> タグ中の<TD> タグの内容を抽出する .
3. 次の式を計算する .

$$Ratio_{real} = \frac{\text{ヒットしたキーワードの重みの総和}}{\text{キーワードの重みの総和}}$$

$$Ratio_{no} = \frac{\text{ヒットしたノイズワードの重みの総和}}{\text{ノイズワードの重みの総和}}$$

4. 得られた値に対して次の式を計算する .

$$Score_i = Ratio_{real} \times \frac{Ratio_{real}}{Ratio_{no}}$$

5. $Score_i$ が閾値 $th1$ 以上であれば , その HTML 文書 d_i に性能表が含まれているとして抽出する .

4.3.4.2 表領域抽出

続いて , 表領域の抽出処理について述べる . 表領域抽出とは , フィルタリングによって抽出された HTML 文書の中から性能表である<TABLE> タグ領域を抽出することを指す . 表領域抽出は , キーワードのみで処理される . 表領域抽出処理の手順を以下に示す .

1. HTML 文書 d_i から<TABLE> タグで記述された領域 $table_j$ を抽出する .
2. <TABLE> タグ中の<TD> タグの内容を抽出し , キーワードが一つも含まれていない場合はその領域を破棄する .
3. $table_j$ に対して , その領域中に存在するキーワードの重みの総和 $Sum_j = \sum_{t \in K} w s_t^{real}$ を計算する . ここで , K はキーワード群である .
4. Sum_j が最大になる領域 $table_j$ について , Sum_j の値が閾値 $th2$ 以上であれば , その領域 $table_j$ を d_i に存在する性能表として抽出する . ここで ,

$$th2 = \frac{\text{キーワードの重みの総和}}{2}$$

である .

表 4.1: エントロピー $M_{real} = 30, M_{no} = 15$

閾値 $th1$	再現率	適合率	F 値
0.20	100.0%	80.7%	91.3%
0.25	100.0%	81.3%	91.6%
0.30	95.0%	82.6%	89.6%
0.35	92.0%	82.1%	87.8%

4.3.5 実験と考察

製品性能表の抽出処理を評価するために5つのサイトから200文書を抽出した。このうち、性能表を含む文書の数100、含まない文書の数100である。また、性能表を含む100文書は性能表だけで構成されているわけではなく、文書中に文字や画像、性能表ではない表も含んでいる。これらの200文書のうち、100文書をキーワード抽出と重み付けアルゴリズムのための訓練データとして用いた。訓練データは性能表を含む50文書と含まない50文書で構成されている。実験はフィルタリングと表領域抽出処理について行った。

まず、HTML文書のフィルタリングについて述べる。評価尺度としては、情報検索などの分野でよく用いられる再現率、適合率および F 値を用いた。

$$\text{再現率}(R) = \frac{\text{正しく抽出されたHTML文書の数}}{\text{性能表を含むHTML文書の数}}$$

$$\text{適合率}(P) = \frac{\text{正しく抽出されたHTML文書の数}}{\text{抽出された全てのHTML文書の数}}$$

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

α は適合率と再現率の相対的な重みを表す。一般的にはこの α を0.5として扱う。これは再現率と適合率を同じに扱うことを意味する。しかし、フィルタリングは、HTML文書から表領域を特定し、性能表を抽出する際の前処理である。つまり、適合率よりも再現率が重視される。そこで、ここでは $\alpha = 0.4$ として、 F 値を計算する。これにより、 F 値は再現率重視の評価値となる。

エントロピーを用いたキーワードとノイズワードおよび重みによる実験結果を表4.1および表4.2に示す。続いて、表4.3および表4.4は、ベイズの定理を用いて作成されたキーワードおよびノイズワードの重みの場合である。表中の M_{real} はキーワードの数、 M_{no} は

表 4.2: エントロピー $M_{real} = 56, M_{no} = 37$

閾値 $th1$	再現率	適合率	F 値
0.20	100.0%	80.0%	90.9%
0.25	100.0%	82.0%	91.9%
0.30	100.0%	85.5%	93.6%
0.35	96.0%	88.1%	92.7%

ノイズワードの数を表す． $th1$ はフィルタリング処理で用いる閾値である．表 4.2 のキーワード数およびノイズワード数は，エントロピーを用いた重み付けによって得られた全キーワードおよびノイズワードの数である．表 4.3 のキーワードは， $P(D_{real}|t) > 0.75$ かつ性能表を含む文書群 D_{real} の半分以上の文書に出現した語，ノイズワードは， $P(D_{no}|t) > 0.75$ かつ性能表を含まない文書群 D_{no} の $\frac{1}{10}$ 以上に出現した語である．表 4.4 のキーワードは， $P(D_{real}|t) > 0.75$ かつ性能表を含む文書群 D_{real} の $\frac{1}{10}$ 以上に出現した語，ノイズワードは， $P(D_{no}|t) > 0.25$ かつ性能表を含まない文書群 D_{no} の $\frac{1}{10}$ 以上に出現した語である．ベイズの定理を用いた場合の W_1 および W_2 は重みの種類を示しており， W_1 は，4.3.3 節で示したベイズの定理で得られた確率そのものを重みとして用いたものである． W_2 はある単語 t のキーワードとしての重みをノイズワードの重みで割ったもの，およびその逆を計算した場合に得られる値を重みとしたものである．すなわち， W_1 においては，

$$ws_t^{real} = P(D_{real}|t) \quad ws_t^{no} = P(D_{no}|t)$$

であり， W_2 においては

$$ws_t^{real} = \frac{P(D_{real}|t)}{P(D_{no}|t)} \quad ws_t^{no} = \frac{P(D_{no}|t)}{P(D_{real}|t)}$$

である．

実験結果により，エントロピーを用いた手法よりもベイズの定理を用いた手法の方が高い再現率および適合率を得た．ベイズの定理を利用した場合， W_2 のように重みの範囲を $0 \sim$ に拡張するよりも $0 \sim 1$ とした W_1 の場合の方が良い結果を得られた．エントロピーによる手法も理論的には重みの範囲は $0 \sim$ であり，キーワードおよびノイズワードの重みはある程度の範囲に正規化された値の方が良い結果を得られることが確認された．エントロピーによる手法では，キーワードおよびノイズワードを全て使った場合の方が若干精度がよいが，有意な差であるとはいえず，キーワードおよびノイズワードを追加する

表 4.3: ベイズの定理 $M_{real} = 19, M_{no} = 15$

閾値 $th1$	重み	再現率	適合率	F 値
0.20	W_1	100.0%	92.6%	96.9%
	W_2	100.0%	90.9%	96.2%
0.30	W_1	100.0%	94.3%	97.7%
	W_2	100.0%	95.2%	98.0%
0.40	W_1	100.0%	98.0%	99.2%
	W_2	99.0%	97.1%	98.2%
0.50	W_1	100.0%	99.0%	99.6%
	W_2	99.0%	98.0%	98.6%
0.60	W_1	98.0%	98.99%	98.4%
	W_2	99.0%	98.0%	98.6%

ことの有意性は低い．ベイズの定理を利用した手法では，むしろキーワードおよびノイズワード数を増加させた方が全体の精度が落ちる傾向がみられた．これにより，提案した手法，特にベイズの定理を用いた手法は少ないキーワードおよびノイズワードで，性能表を含んだ HTML 文書のフィルタリングが行えることが確認された．

続いて，表領域抽出処理について述べる．フィルタリングによって得られた結果を用いて表領域を抽出した．エントロピーを用いた手法では， $M_{real} = 30$ および $M_{no} = 15$ でもっともフィルタリングの精度が良かった閾値 $th1 = 0.25$ の実験結果を用いた．同様にベイズの定理を用いた手法では， $M_{real} = 19$ および $M_{no} = 15$ で閾値 $th1 = 0.50$ の実験結果を用いた．表領域の特定処理においても，ベイズの定理を用いた手法の方が，高い再現率および適合率を得た．エントロピーを用いた手法およびベイズの定理を用いた手法の双方とも適合率に比べ，再現率が若干低い．これは表領域抽出処理で用いた閾値 $th2$ の決め方に問題があるためだと考えられる．この閾値 $th2$ の決定法は今後の課題の一つである．

総じて，フィルタリングにおいても，表領域抽出処理においても，高い再現率および適合率を得ることができた．性能表抽出処理で用いるキーワードおよびノイズワードの追加は精度向上には有効ではなく，少ないキーワードおよびノイズワードで実現できることが確認された．ベイズの定理においては，フィルタリングでほぼ 100% の F 値 を得た．実験により，提案した手法の有効性が確認された．

表 4.4: ベイズの定理 $M_{real} = 66, M_{no} = 41$

閾値 $th1$	重み	再現率	適合率	F 値
0.20	W_1	100.0%	97.1%	98.8%
	W_2	100.0%	95.2%	98.0%
0.30	W_1	92.0%	97.9%	94.3%
	W_2	92.0%	96.8%	93.9%
0.40	W_1	83.0%	97.65%	88.3%
	W_2	86.0%	98.9%	90.7%
0.50	W_1	76.0%	100.0%	84.1%
	W_2	78.0%	98.7%	85.2%
0.60	W_1	57.0%	100.0%	68.8%
	W_2	63.0%	100.0%	73.9%

表 4.5: 表領域抽出処理 - エントロピー -

閾値 $th1$	再現率	適合率
0.25	93.0%	96.9%

4.3.6 他研究との比較

関連研究の節でも述べたが，Web 上の情報の有効利用という観点から，表抽出や解析に関する研究が頻繁に行われるようになってきている．我々と同様に，ある特定のドメイン（航空会社の Web サイト）を対象を絞り，人で作成されたヒューリスティックによるルールを用いて表の抽出を行った Chen ら [5] の手法は，再現率 = 80.1%，適合率 = 92.9% である．ルールやキーワードをドメインごとに作り直すコストの点から考えても，提案した手法は，Chen らの手法より有効である．

Wang ら [73] は，<TABLE> タグで記述された領域が本当に表であるかどうかを判定する手法として，決定木アルゴリズムと Support Vector Machines (SVM) を比較・考察している．Wang らは，<TABLE> が表であるかを判別する特徴として，文字データだけでなく，レイアウト的な特徴や表を構成する要素に用いられているデータ形式の種類（例えば，画像があるか，ハイパーリンクがあるかなど）も考慮している．これらは，Wang

表 4.6: 表領域抽出処理 - ベイズの定理 -

閾値 $th1$	再現率	適合率
0.50	95.0%	100.0%

らの目的のように、表かどうかを判定するには有効だが、本研究で対象としているような、「ある特定の内容が記述されている表のみを抽出する」というタスクでは、あまり重要ではない。直観的に考えても、レイアウト情報から「性能表」と「周辺機器一覧」のような見た目が酷似した表を分類することは困難である。そこで、今回は Wang らの提案した手法のキーワードの重み付け部分のみを対象とし、精度を比較する。Wang らは、表である<TABLE> とそうでない<TABLE> について正規化された $tf \cdot idf$ 値を求め、それらの比率を属性値として決定木を作成している。得られた決定木を用いてフィルタリング処理を行なったところ、再現率 = 90.0%、適合率 = 93.8%、 F 値 = 91.5% となった。実験結果と比較して、提案した手法の方が有効であることが確認された。また、Chen らや Wang らは、表データの有用性を踏まえ、表の抽出を行っているが、具体的な利用方法について、十分に議論していない。我々が、これらの研究と大きく異なるのは、表から得られたデータを扱うことの有効性を検証していることにある。詳しくは 4.5 節以降で述べる。

Yoshida ら [82] は、EM アルゴリズムによる表認識手法について述べている。Yoshida らの目的は、表から抽出したデータを用いてオントロジーの構築するというものであり、表データの有用性について議論している。しかし、この手法の再現率は 85.9%、適合率は 79.4% であり、十分な精度が得られているとはいえない。

4.4 製品性能表の解析

処理対象としている製品性能表は HTML で記述された 2 次元の表とする。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により先頭の行や列だけにそのデータが記述されている。図 4.5 (a) の表の HTML 表記が図 4.5 (b) である。一般に表の 1 列目は製品の属性で、残りの列には各製品に対するデータで構成されている。

これら表における問題点は、その表記が各メーカー・製品ごとに異なり、正規化されていないことである。具体的には (1) 属性名が正規化されていない、(2) 1 つのセルに複

数のデータが存在する，(3) 幾つかの属性はサブカテゴリを持つ（例：図 4.5 (a) の “Memory”），(4) 幾つかのデータは共有される（例：図 4.5 (a) の “256MB”），が挙げられる．

4.4.1 表構造の定義

これらの問題を解決するために，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．

(列項目 行項目 データ)

ここで，列項目にはモデル名，行項目には属性名，データにはその属性に対するデータが割り当てられる．殆どの性能表では，行項目は表の上部に存在し，列項目は表の左側に存在する．列項目および行項目はしばしば複数存在し，その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_m) (行項目₁... 行項目_n) データ)

表解析のための知識として，(1) 属性の標準表記，(2) 属性と表中に出現する単位との関係，(3) 専門的な単語群，をあらかじめ人手で定義してある．

4.4.2 表構造の生成

表構造生成のアルゴリズムを以下に示す．

1. HTML における複数の列や行を記述するためのタグ (colspan および rowspan) を基に表の属性とデータを対応付ける．

図 4.6 は図 4.5 (a) について属性とデータの対応付け処理を行った結果である．図中の破線部分は補完された線である．

2. 対応付けられた表を表構造へ変換する．

ここで， $c(i, j)$ を i 行目・ j 列目のセル， $(列項目_k 行項目_k データ_k)_k$ を k 番目の表構造，表のサイズが $M \times N$ だと仮定する．

(2-1) 表中の “製品名” や “モデル名” などのキーワードと位置情報から表構造の列項目 _{k} にモデル名を割り当てる．この時の i を b_i とする．

(2-2) 行項目の幅を特定する．

行項目の幅は以下の式から特定される．

$$Score(C_j) = \sum_i^M B_{ij} \quad (j = 1, \dots, N - 1)$$

Model Name		PC1	PC2
CPU		400MHz	450MHz
Memory	Std	64MB	128MB
	Max	256MB	
	VRAM	4MB	

(a)

```

<table border="1">
  <tr>
    <td colspan="2"> Model Name</td>
    <td>PC1</td>
    <td>PC2</td>
  </tr>
  <tr>
    <td colspan="2">CPU</td>
    <td>400MHz</td>
    <td>450MHz</td>
  </tr>
  <tr>
    <td rowspan="3">Memory</td>
    <td>Std</td>
    <td>64MB</td>
    <td>128MB</td>
  </tr>
  <tr>
    <td>Max</td>
    <td colspan="2">256MB</td>
  </tr>
  <tr>
    <td>VRAM</td>
    <td colspan="2">4MB</td>
  </tr>
</table>

```

(b)

図 4.5: HTML 表記による表の例

Model Name		PC1	PC2
CPU		400MHz	450MHz
Memory	Std	64MB	128MB
Memory	Max	256MB	256MB
Memory	VRAM	4MB	4MB

図 4.6: 対応付けられた表

ここで, $Score(C_j)$ は j 列目の得点, B_{ij} は i 行目・ j 列目に境界があれば 1, なければ 0 となる. この $Score(C_j)$ が最大となる列が行項目の幅となる. 例を図 4.7 に示す. この時の j を b_j とする.

(2-3) $i = b_i + 1, \dots, M$ において,

- (1) 求められた幅から行項目 k に属性を割り当てる.
- (2) $c(i, j)$ をデータ k とする. ($j = b_j + 1, \dots, N$)
- (3) $k = k + 1$

3. 得られた表構造に対して以下の処理を行う.

(3-1) 表構造中の語の表記を統一する. 人手で作られた辞書を用いる. 辞書に登録されている語数は 50 語である.

例: モニター, スクリーン \Rightarrow ディスプレイ

(3-2) 行項目が“数値 + 単位”を含んでいれば, 行項目とデータを変更する.

例: (PC1 (解像度 1024 × 768dpi)) \Rightarrow
(PC1 解像度 1024 × 768dpi)

(3-3) データに複数値が存在する場合, “,” や “/” などの記号から, それらを複数の表構造に分割する

例: (PC1 インタフェース (USB × 2, IEEE × 1)) \Rightarrow
(PC1 インタフェース USB × 2)
(PC1 インタフェース IEEE × 1)

(3-4) データが特定の語・記号・数値 (例: USB × 2) を含んでいれば, 行項目とデータを変更する.

例: (PC1 インタフェース USB × 2) \Rightarrow
(PC1 (インタフェース USB) 2)

(3-5) “[]” や “-” などの表特有の表現があれば, それを正規化する.

例: (PC1 ベイ [空き] 5[2]) \Rightarrow
(PC1 ベイ 5)
(PC1 (ベイ 空き) 2)

図 4.8 は図 4.6 の表から生成された表構造を表す.

4.4.3 実験と考察

表構造生成処理の実験結果について述べる. 評価には上で述べた性能表を含む 100 文書 (製品数は 247 個) を用いた. 表 4.7 にその実験結果を示す. ここで, 表構造が正しく生成されたとは, 列項目であるモデル名, 行項目である各属性およびデータ部が認識され, 正しく表構造に変換されたことを指す. 生成に失敗した原因は HTML 中のタグの不

C1=3 C2= 5 C3=3

Model Name	PC1	PC2
CPU	400MHz	450MHz
Memory	Std	128MB
	Max	256MB
	VRAM	4MB

図 4.7: 行項目の特定

(PC1 CPU 400MHz)
 (PC1 (Memory Std) 64MB)
 (PC1 (Memory Max) 256MB)
 (PC1 (Memory VRAM) 4MB)
 (PC2 CPU 450MHz)
 (PC2 (Memory Std) 128MB)
 (PC2 (Memory Max) 256MB)
 (PC2 (Memory VRAM) 4MB)

図 4.8: 表構造の例

表 4.7: 表構造生成処理の実験結果

	精度
サイト A	100%(10/10)
サイト B	85%(34/40)
サイト C	100%(30/30)
サイト D	100%(10/10)
サイト E	100%(10/10)
計	94%(94/100)

足である。一般の Web ブラウザは、HTML の各セルの終了を表す</TD> タグや各行の終了を表す</TR> タグがない場合でも、曖昧性が生まれえない場合はそれらのタグが存在するものとして表示する。現在の我々の表構造生成プログラムはこれらのタグ情報を用いて解析しているため、タグの欠落により正しい表構造が得られなかった。これらのタグの補完が行えれば、表構造生成処理の精度は 100% に達する。

4.5 特徴データ抽出処理

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を表したデータなのかはわからない。本節では、各々の製品を特徴づけるような属性やデータ（以下、特徴データと呼ぶ）を抽出し、それらをユーザの要求に併せてスコアリングする手法について述べる。表中の属性は定量的項目と定性的項目の 2 つに分類される。表 4.8 にその分類例を示す。

4.5.1 定量的項目の特徴データ抽出

まず、定量的項目に関する特徴抽出法について述べる。特徴データは、表中の単位の前にある数値を比較することで抽出される。用いられる単位としては、“MHz”、“MB”、“色”、“mm”、“W” などがある。これらの単位をその直前の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た。分類結果を表 4.9 に示す。この分類を基に数値の比較を行う。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データから標準値を求め、それを基準に比べる場合の 2 種類がある。すべての製品が同じ値を持つ属性は特徴データとして抽出しない。比較処理の前処理として、次の処理を実

表 4.8: 定量的項目と定性的項目

定量的項目	定性的項目
CPU クロック: MHz, GHz	CPU プロセッサの種類
メモリ: MB	グラフィックチップ
ディスプレイ: インチ	デバイスの種類
重量: kg	(CD-R/RW や DVD-ROM, DVD-RAM など)
サイズ: mm, cm	OS の種類
...	...

表 4.9: 比較のための単位の分類

数値が大きいほどよい単位	MHz, KB, MB, GB, 色, ドット, 型, インチなど
数値が小さいほどよい単位	W, 円, Kg など
属性によって変わる単位	時間, mm など

行する .

- 値の正規化: (mm \Leftrightarrow cm) (MHz \Leftrightarrow GHz) (KB \Leftrightarrow MB \Leftrightarrow GB)
- 範囲があるデータはその最小値 (もしくは最大値) に変更する
(PC1 サイズ 38-40mm) \Rightarrow (PC1 サイズ 38mm)

標準値の算出方法は、単位により異なり、“kg” や “時間” など直前の数値が連続的に増減するものは、全データの平均値を標準値とする。また、“MB” や “色” のように直前の数値の増え方に一定のパターンがあるものは、最頻値の数値を標準値とする。それぞれの算出方法に分類される単位の一覧を表 4.10 に示す。得られた標準値を用いて比較を行い、その結果が標準値より大きい場合もしくは小さい場合は、数値が表構造中での最大値もしくは最小値でなくとも、それを特徴データとして抽出する。

具体的には、最大値を 10 ポイント、標準値を 5 ポイント、最小値を 0 ポイントとし、すべてのデータに対してスコアリングを行う。最大値・標準値間の 1 ポイントごとの値と、最小値・標準値間の 1 ポイントごとの値を計算し、その値を基に各製品の属性にスコアがつけられる。図 4.9 に例を示す。ここで、すべての製品データから最大値：1.1GHz，

表 4.10: 標準値の算出処理のための分類

平均値	W, 円, mm, 時間, Kg など
最頻値	MHz, KB, MB, GB, 色, ドット, 型, インチなど

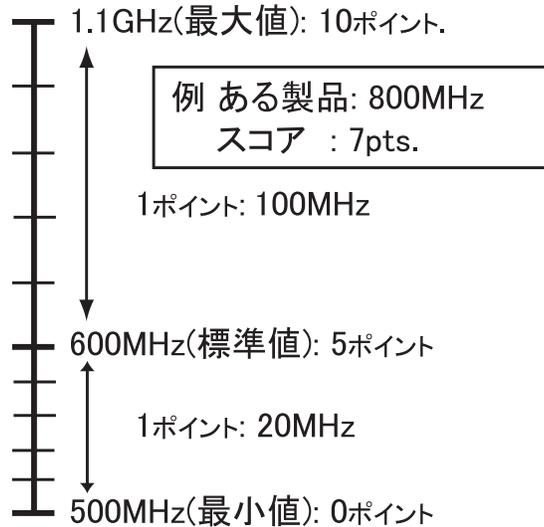


図 4.9: スコアリング

標準値：600MHz，最小値：500MHz が得られていると仮定する．もし，ある製品のデータが800MHzだと仮定すると，その製品のスコアは7ポイントとなる．

製品性能表は大部分が定量的項目に属する数値データで構成されており，表 4.9および表 4.10を作成することで殆どの特徴データを抽出することができる．

4.5.2 定性的項目の特徴データ抽出

続いて，定性的項目についての特徴抽出法について述べる．定性的項目に関する特徴抽出法は人手によって作成された辞書を用いた特徴データ抽出と関連項目を用いた特徴データ抽出の2つに分けられる．まず，辞書を使った特徴データの抽出法である．この辞書には特徴データになる可能性があると考えられるものをあらかじめ登録しておく．現在登録されている単語数は23個であり，各単語は重みを持っている．これらの単語が表構造中にあれば，そのデータをその製品の特徴データとして抽出し，スコア付けする．処理の流れを以下に示す．

関連項目：VRAM ⇔ グラフィックボード 仮定：PC3のVRAMが既に特徴データとして抽出されている。

(PC3 グラフィックボード GB1000PRO)
 ↓
 特徴データ：GB1000PRO

図 4.10: 関連項目による特徴データ抽出の例

1. 表構造中で辞書に登録されている単語を探索。
2. 単語が表構造中にある場合
 - (2-1) 比較対象である他のモデルの表構造中にその単語があるかを調べ、すべてのモデルにある場合は抽出しない。
 - (2-2) 辞書中の重みに基づきスコアづけを行う。
 例：(Pentium4: 4, Pentium3: 3, Celeron: 2)

辞書に登録されている単語は製品性能表で一般に使われる単語であり、搭載されているグラフィックボードなどの固有名詞やそれらに関する情報すべてを辞書に登録しておくことは現実的に困難である。そこで、4.5.1節で抽出された特徴データと関連項目を用いて特徴データを抽出する。関連項目とは、あらかじめ人手により作成された表の属性間の関係リストであり、その数は現在6個である。関連項目を用いた特徴データ抽出処理では最大値および標準値によって抽出された特徴データの両方を用いる。まず、最大値として抽出された属性の関連項目を特徴データとして抽出する。続いて、標準値によって抽出された属性の関連項目も特徴データとして抽出する。

図 4.10は、関連項目を用いた特徴データ抽出の例である。表の属性“VRAM”と“グラフィックボード”は関連項目として定義されている。これはあるモデルの“VRAM”の値が特徴データとして抽出されていれば、そのモデルの“グラフィックボード”も特徴データとして抽出することを表している。4.5.1節の処理でPC3の“VRAM”の値が既に特徴データとして抽出されていると仮定すると、PC3の“グラフィックボード”である“GB-1000PRO”も特徴データとして抽出される。

表 4.11: ユーザからの要求と関連する属性

ユーザからの要求	属性
高性能	CPU, メモリ, HDD, ディスプレイ, ...
グラフィックスに強い	グラフィックチップ, CPU, メモリ, ...
モバイル PC	重量, 本体サイズ, バッテリー駆動時間, ...
実用性重視	CPU, HDD, 価格, ソフト, ...
価格重視	価格, ソフト, CPU, メモリ, ...

4.5.3 ユーザの要求に応じたスコアリング

前節までで述べた定量的項目および定性的項目を用いた手法は、静的な特徴データ抽出である。ユーザが求めている製品を効率よく見つけだすために、我々は表中の属性とユーザの要求の対応関係を定義した。表 4.11 にその対応関係を示す。現在我々のプロトタイプシステムではユーザからの 5 つの要求を扱える。

さらに、ある程度知識があるユーザもしくは選択の際に重要な属性がある程度わかっているユーザは、スコアリングの際に用いる表中の属性の重みを変更することができる。図 4.11 に選択用のウィンドウを示す。ユーザは -1 から 4 までの範囲で重みを指定できる。また、属性同士にはある程度関連性がある。例えば、属性“重量”は携帯性という観点から“本体サイズ”や“バッテリー駆動時間”などと関連がある。表 4.12 に関連性がある属性の例を示す。この関連性を用いて、図 4.11 中の上部にある“Related Items”をクリックすることで、プロトタイプシステムは重みを自動的に拡張する。属性間の関連性を用いることで、製品に対して十分な知識がないユーザにとって、より柔軟な選択ができるようになる。関連属性の重みは次のように設定される。

$$uw(j) = \begin{cases} 0.5 & (uw(i) = 1) \\ -0.5 & (uw(i) = -1) \\ uw(i) - 1 & (\text{The others}) \end{cases}$$

ここで、 i はユーザが選んだ表中の属性、 j は属性間の関連性によって拡張された属性を示す。 $uw(i)$ はユーザが指定した重み、 $uw(j)$ は拡張された属性の重みである。例えば、ユーザが属性“重量”の重みを 2 と設定し、“Related Items”をクリックすると、属性“本体サイズ”や属性“バッテリー駆動時間”の重みが自動的に 1 となる。

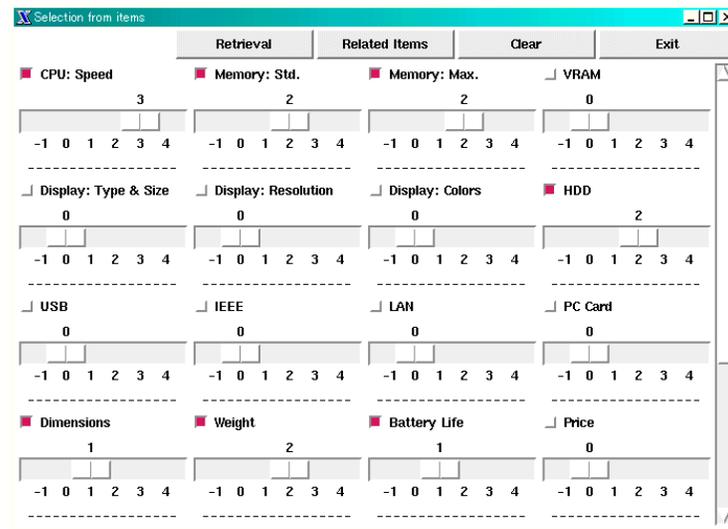


図 4.11: 重み選択ウィンドウ

表 4.12: 属性間の関連性

属性	関連属性
CPU	メモリ（標準・最大），HDD
ディスプレイ	VRAM，解像度，色
USB	IEEE1394，PC カード
重量	本体サイズ，バッテリー駆動時間

続いてスコアリングの流れを以下に示す。

1. ユーザの要求に関連する属性を伴う表構造を探索する。
2. 各製品に対して以下の式でスコアリングを行う。

$$score(c, r) = \frac{\sum_{k=1}^n (w(a_k, r) + uw(a_k)) \times pt(a_k, c)}{\sum_{k=1}^n w(a_k, r)}$$

ここで、 c 、 r 、 a_k はそれぞれ製品、ユーザの要求、属性を表す。 $w(a_k, r)$ は各要求 r における属性 a_k の重みである。 $w(a_k, r)$ は事前に定義してある。 $pt(a_k, c)$ は前節までで計算されているスコアである。

3. スコアが閾値以下の製品を削除する。
4. スコアに基づき降順に出力する。

4.5.4 適合性フィードバック

一度の検索で、ユーザが必要とする情報がすべて得られるとは限らない。これはスコアリングに用いた属性の重みが静的なためである。そこで、ユーザに出力結果を提示し、ユーザがその結果を見て、システムの重みを変化させるようにする必要がある。このような技術を一般に適合性フィードバックと呼ぶ。

プロトタイプシステムでは、結果は既に順序づけされており、ユーザは上位の幾つかの結果を選んで適合性判断を行うことができる。適合性判断された結果を用いて再度スコアリングを行う。ここで各属性の重みを検索質問だと仮定すると、初期入力質問は $Q_0 = (w(a_1, r), \dots, w(a_n, r))$ と見なすことができる。これにより、修正された検索質問は Q_1 であり、

$$Q_1 = Q_0 + \alpha \sum_{i=1}^{N^+} D_i^+ - \beta \sum_{i=1}^{N^-} D_i^-$$

によって、ユーザの意図を反映した属性の重みが生成される。ここで、 N^+ および N^- はユーザが選択した適合文書および不適合文書の数であり、 D^+ および D^- はそれぞれ適合文書、不適合文書における重みのベクトルである。 α および β は定数であり、それぞれ適合文書、不適合文書をどの程度重要視するかを調整するために用いられる。

4.5.5 実験と考察

(1) 特徴データ抽出処理、(2) ユーザの要求と適合性フィードバックについて実験を行った。

4.5.5.1 特徴データ抽出処理実験

まず、製品性能表からの特徴データ抽出処理について考察する。47 機種の商品性能表を用い、極端に性能差がある機種同士を比較しないように、それらを7つに分けて実験を行った。実験に使用した製品性能表の例を図4.12に示す。また図の製品性能表から抽出された特徴データの一覧を表4.13に示す。表中のMおよびSの記号はMが最大値もしくは最小値によって特徴データとして抽出されたもの、Sは標準値によって特徴データとして抽出されたものを表している。

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。雑誌による評価を要約すると以下のようになった。

PC1 の特徴

- (1) 基本性能が比較的良好い。
- (2) ビデオ編集などに適している。
- (3) 若干サイズが大きい。

PC2 の特徴

- (4) 大容量のハードディスクを搭載。
- (5) 表示能力が高い。

PC3 の特徴

- (6) 基本性能が高い。
- (7) 表示能力が比較的高い。
- (8) ビジネスでの使い勝手も良い。

このレビューを基に実験結果を考察するとPC1ではメモリが最大値として抽出されており、CPUとHDDも標準値として抽出されている。これらはPC1の特徴(1)から考えて妥当な特徴データだといえる。またIEEE1394が特徴データとして抽出されている点も特徴(2)を考えれば妥当である。続いてPC2ではHDDが最大値で抽出されており、これは特徴(4)から正しい特徴データだといえる。ディスプレイやVRAMが抽出されていることも特徴(5)から妥当だといえる。最後にPC3についてはCPUおよびメモリが最大値で抽出されており、これらは特徴(6)から正しい特徴データだといえる。VRAMの値が抽出されていることは特徴(7)から、LANを装備していることは特徴(8)から妥当な結果である。サイズや重量などが抽出されることは持ち運びに便利なことを意味しており、これらも特徴(8)を考えれば妥当な特徴データだといえる。以上のように大部分のデータがレビューと比較して正しく特徴データを抽出できているといえる。またこれ以外のデータでも同様にレビューとの比較を行った結果、同等の妥当性を得ることができた。

また、レビューとの比較とは別に、特徴データを人手によって評価した。被験者は本研究に直接関係していない大学院生8名である。実験データは先ほどの実験データとは別の製品性能表22機種分を5つに分けたものである。被験者に実験で用いた製品性能表とそ

		PC1	PC2	PC3
CPU		650MHz	600MHz	700MHz
メモリ	標準	128MB	64MB	128MB
	最大	256MB	256MB	512MB
HDD		18GB	20GB	12GB
ドライブ		内蔵 DVD-ROM6 倍速	内蔵 DVD-ROM6 倍速	内蔵 DVD-ROM6 倍速
ディスプレイ		14.1 インチ TFT 液晶	15 インチ TFT 液晶	14.1 インチ TFT 液晶
VRAM		6MB	8MB	8MB
解像度		1024 × 768 ドット	1024 × 768 ドット	1024 × 768 ドット
表示色		1677 万色	1677 万色	1677 万色
通信機器		56k モデム	56k モデム, 携帯 / PHS	56k モデム, LAN
PC カード		TYPE II × 2 または TYPE III × 1	TYPE II × 2 または TYPE III × 1	TYPE II × 2 または TYPE III × 1
インタフェース		USB × 1, IEEE1394	USB × 3, IEEE1394	USB × 1
バッテリー時間		3 時間	2.1 時間	4 時間
サイズ		308mm(W) × 261mm(D) × 45mm(H)	337mm(W) × 289mm(D) × 41.3mm(H)	304mm(W) × 250mm(D) × 33.1mm(H)
重量		3.1kg	4.2kg	2.36kg

図 4.12: 実験に用いた製品性能表の例

表 4.13: 抽出された特徴データ

	属性	データ	種類
PC1	CPU	650MHz	S
	メモリ 標準	128MB	M
	HDD	18GB	S
	インタフェース	IEEE1394	M
	重量	3.1kg	S
PC2	HDD	20GB	M
	ディスプレイ	15 インチ TFT 液晶	M
	VRAM	8MB	M
	通信機器	携帯 / PHS	M
	インタフェース	USB × 3, IEEE1394	M
PC3	CPU	700MHz	M
	メモリ 標準	128MB	M
	メモリ 最大	512MB	M
	VRAM	8MB	M
	通信機器	LAN	M
	バッテリー時間	4 時間	M
	サイズ	304mm(W) × 250mm(D) × 33.1mm(H)	M
	重量	2.36kg	M

表 4.14: 実験結果

	再現率	適合率
データ 1 (4 機種)	100%	87.0%
データ 2 (7 機種)	100%	89.3%
データ 3 (5 機種)	100%	96.4%
データ 4 (3 機種)	100%	91.2%
データ 5 (3 機種)	100%	97.1%

の製品性能表から抽出された特徴データのリストを渡し、特徴データが妥当であるか、抽出漏れがないかをチェックした。評価尺度には、表抽出・解析と同様に再現率と適合率を用いた。実験の結果を表 4.14 に示す。抽出結果の再現率からすべての特徴データを抽出できたことがわかる。適合率が再現率に劣る理由として、本手法は標準値による特徴データ抽出の際、そのデータが標準値より大きいかもしれないか小さいかだけで判断するため、被験者が標準値に近い値で特徴データとして抽出されているものを適切でないと判断したことなどが挙げられる。しかし、実際にはユーザの要求により重み付けすること、適合性フィードバックによってユーザ検索意図を反映できること、要約処理で文章と一緒に表も出力することなどから考えれば、適合率が 90% 前後であれば問題はないと考えられる。これにより、本手法の有効性が確認された。

4.5.5.2 ユーザの要求と適合性フィードバックを用いた検索

我々はプロトタイプシステムを用いて、ユーザの要求と適合性フィードバックについて評価した。評価にはノート型 PC の性能表を用い、製品数は 38 製品である。現在のプロトタイプシステムが扱える要求は、(1) 高性能、(2) グラフィックスに強い、(3) モバイル PC、(4) 実用性重視、(5) 価格重視、の 5 種類である。我々はシステムの出力結果と雑誌“日経ベスト PC”[40] で推奨された製品との比較を行い、本システムの出力結果の妥当性を評価した。表 4.15 は我々のシステムがスコアによって 1 位にランク付けした製品の雑誌中での順位を示している。また表 4.16 は雑誌中で 1 位にランク付けされた製品が我々のシステムで何位にランク付けされているかを示している。要求 (3) および (5) に関しては、システムと雑誌間で 1 位にランク付けされた製品が同じものであり、またそれ以外の上位にランク付けされた製品もシステムと雑誌間で一致した。要求 (1) および (4) に関しても、システムと雑誌間での順位は似通っており、妥当なスコアリン

表 4.15: システムによって 1 位にランクされた製品の雑誌中での順位

ユーザの要求	雑誌中での順位
要求 (1)	4 位
要求 (2)	8 位
要求 (3)	1 位
要求 (4)	4 位
要求 (5)	1 位

表 4.16: 雑誌中で 1 位にランクされた製品のシステム中での順位

ユーザの要求	システム中での順位
要求 (1)	3 位
要求 (2)	5 位
要求 (3)	1 位
要求 (4)	4 位
要求 (5)	1 位

グが行えているといえる。要求 (2) については、幾つかの製品はシステムと雑誌間の順位が一致しないが、雑誌中で推奨された 2 つの製品が、システムの上位 5 位までに含まれており、スコアリングが妥当であったことが確認された。

続いて、適合性フィードバックについて検証する。適合性フィードバックでは、式中の α および β をそれぞれ $\frac{1}{N^+}$, $\frac{1}{N^-}$ とした。この式は Rocchio の式としてよく知られている [50]。これは、すなわち、重みの調整分を適合、不適合文書数で正規化したものである。図 4.13 は適合性フィードバックを用いた場合の適合率・再現率のグラフである。これは、あるユーザが要求として“モバイル PC”を選択したケースである。ここで、そのユーザはとりわけ携帯性を重要視していた。すなわち、“CPU”や“メモリ”などの属性よりも“重量”や“本体サイズ”、“バッテリー駆動時間”などの属性を重要視していた。ユーザは得られた結果に対し、上位 5 製品について適合性判断を行った。その適合性判断から Q_1 を再計算し、それらを用いて再度検索を行った結果、適合率および再現率が大幅に上昇し、本システムの有効性が確認された。図 4.13 中の“ Q_0 with user's weights”は 4.5.3 節で述べたユーザによる属性の重み設定を行った場合の結果である。ここで、ユーザは属性

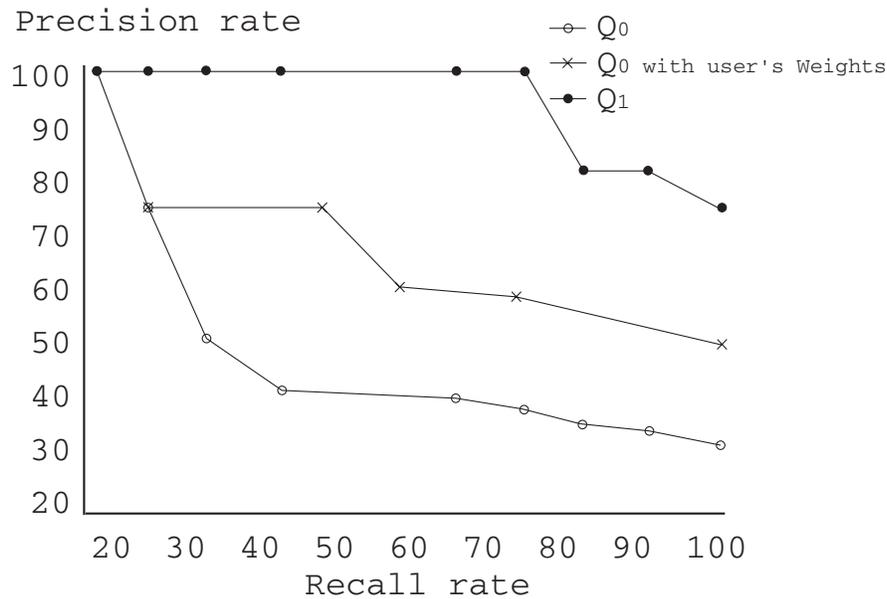


図 4.13: 適合率と再現率

“重量”の重みを3とし，“Related Items”ボタンをクリックした。

4.6 出力形式

前節では，製品の特徴データの抽出とスコアリング手法について述べた．前節で求められたスコアを用いれば，その結果を順序づけてユーザに提示することが可能である．しかしながら，特徴抽出に用いた表のままをユーザに情報を提示することは，わかりやすい出力形式であるとは言えない．また，複数の表を見比べるには労力がかかる．本節では，前節で特徴データとして抽出，スコアリングされたデータをよりわかりやすい形で提示するために手法について述べる．

4.6.1 文章による要約生成

プロトタイプシステムでは，ユーザの選択した製品群について，それらの特徴データおよびスコアを用いて文章を生成することができる．生成される文書の構造を図 4.14に示す．文書は，どんな製品についての要約文かを表す大見出しと，あらかじめ定義された話題に関連した小見出し i および本文 i の集合で構成される．

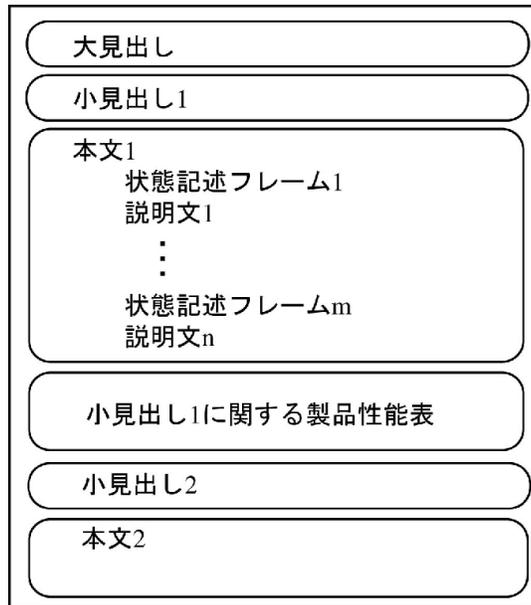


図 4.14: 文書の構造

4.6.1.1 文章生成フレームと説明文

文章は、あらかじめ用意された文章生成フレームのスロットに抽出された特徴データをセットすることで生成される文と、生成された文章で用いられた表の属性もしくはデータによって生成される説明文の2種類に分けられる。特徴データをスロットにセットすることにより生成される文章生成フレームは3種類ある。説明文にはスロットはない。例を以下に示す。

- 大見出しフレーム
 - 「社名」は{「モデル名」、「シリーズ名」、「モデル数」シリーズ}を発表した。
- 小見出しフレーム
 - 「話題」の{優れた, 高性能な, よい}「モデル名」。
- 状態記述フレーム
 - 「モデル名」は「属性」に「データ」を{搭載, 採用, 装備}。
- 説明文
 - 条件: 「属性」= グラフィックチップ, グラフィックボード
 - 説明文: 高性能のグラフィックチップの搭載により画像処理を快適に行える。

フレーム中の「モデル名」や「シリーズ名」には表構造の列項目，「属性」には行項目，「データ」にはデータがセットされる．「話題」は，あらかじめ与えられており，現在9個ある．「話題」は文章中の話題にあたる．「話題」とそれに関連する「属性」や「データ」の一覧を表4.17に示す．小見出しの「モデル名」（以下，話題モデルと呼ぶ）は，以下の式で決定される．

$$TM = \operatorname{argmax} \sum_{a_i \in Topic} \operatorname{Score}(C_j, a_i)$$

ここで， TM は話題モデル， a_i は属性， C_j は製品， $Topic$ は現在の話題， $\operatorname{Score}(C_j, a_i)$ は製品 C の a_i のスコアである．すなわち，特徴データ抽出処理で得られた各属性のスコアの総和が最大となるモデルがセットされる．

フレーム中の「属性」および「データ」には，表から抽出された特徴データが，そのフレームにセット可能かどうかを判断するため，セット可能な属性およびデータのリストが制約としてセットされている．この制約に基づき，表データがフレームに割り当てられる．現在の文章生成フレームで扱う制約を以下に示す．

- 「AllA」：全ての属性（表構造の行項目）
- 「AllV」：全てのデータ（表構造のデータ）
- 「SA」：指定された属性のみセット可能
- 「SV」：指定されたデータのみセット可能

具体的な使用例を以下に示す．

（例）「モデル名」は「SA-:USB|Serial|IEEE」ポートを「SV-:[0-9]+」基装備している．

この例は，2番目のスロットに表の属性であるUSB，SerialもしくはIEEEのどれかがセット可能であり，かつ3番目のスロットにデータの数値部分がセット可能な文章生成フレームであることを示している．

4.6.1.2 文章生成処理

文章生成処理の流れを以下に示す．

1. 特徴データを含む表構造の抽出．

ここで，特徴データを含む表構造とは，特徴データ抽出処理でスコアが5以上のものを指す．

表 4.17: 話題と対応する属性

話題	関連する表の属性とデータ
基本性能	CPU, メモリ, ハードディスクなど
拡張性	PCI, USB, PC カードなど
画像処理	グラフィックボード, 動画処理ソフトなど
表示機能	ディスプレイサイズ, VRAM など
操作性	キーピッチ, キーストロークなど
携帯性	サイズ, 重量, バッテリー駆動時間など
通信	モデム, LAN 端子など
音楽	スピーカ, サウンドボードなど
ソフト	OS, 添付ソフトなど

2. 話題ごとに表構造を分類 .
3. モデルごとに表構造を分類 .
4. 製品性能表中での提示順序に基づき, 表構造をソート .
話題モデルの表構造を優先的に出力 .
5. 条件により, 説明文を出力 .

文章として出力される際に特徴データ抽出処理で得られたスコアに基づき, 文章生成フレームが選択される . 具体的にはスコアが 10 ポイントのものは, そのまま文章生成フレームが適用され, それ以外の場合は, スコアに応じて “ 比較的 ” などの語を補ったフレームが適用される . 処理の例を図 4.15 に示す .

4.6.1.3 修辞処理

前節で特徴データをセットした文章生成フレームをそのまま出力すると文章の羅列となり, 不自然である . そこで, 必要に応じて, 語の省略, 接続詞の補完などを行う . これらの条件と処理の例を以下に示す .

- 文章の統合

条件 連続する 2 つの文章生成フレームにセットされているモデル名が異なり, 表の属性が同じである .

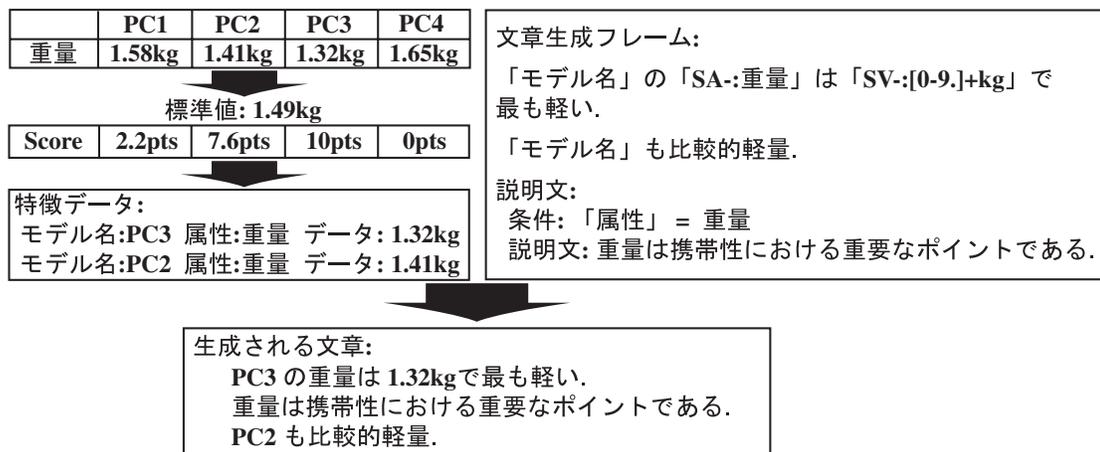


図 4.15: 文章生成処理の例

処理 2つの文章を1文に統合.

- モデル名の削除

条件 連続する2つの文章生成フレームにセットされているモデル名が同じである.

処理 後ろの文章のモデル名を削除.

2つの文章の間に「また」、「同様に」などの接続詞を挿入.

これらの修辞処理は複数同時に適用されることもある.

4.6.2 表の再構成

表を文章として要約すると、重要なデータは抽出され、文章として生成されるものの、それが他の製品のデータとどの程度違うのかはわからない。具体的な性能差に関する情報は文章にすると冗長になるため、表同士を見比べた方がわかりやすい。また、文章のみもしくは表のみの出力よりもそれらを統合した要約の方が、読み手にとってわかりやすい。

文章による要約と表を統合するために、話題に関連する部分だけを表構造から抽出し、その表構造を基に表構造から製品性能表への再変換を行う。再変換の手順は表の解析処理を逆にたどることになる。具体的には次のような手順で行われる。

1. 対象となる表構造を抽出.
2. 属性が同じ要素を統合.
3. 属性、データから colspan および rowspan タグを推定.

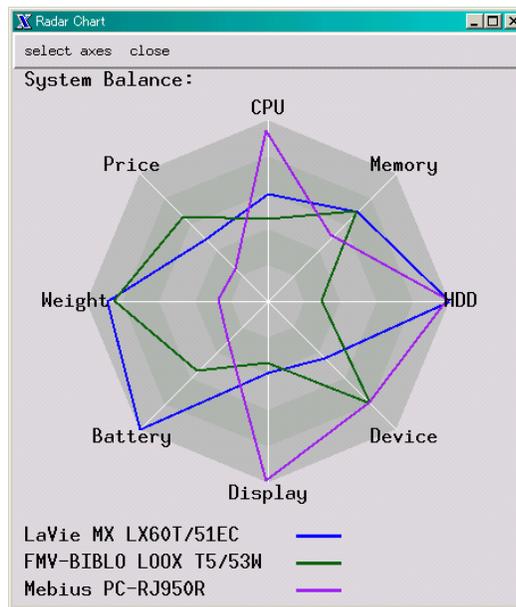


図 4.16: レーダチャート

以上のような手順で再構築された表を文章内の適切な位置に配置して、HTML 形式で出力する。

4.6.3 レーダチャートの生成

文章や表といったデータは、細部まで知りたい場合などには有効だが、一目見て製品間の特徴を知りたい場合には向かない。データを容易に比較するためには、グラフによる出力が有効である。

プロトタイプシステムでは、ユーザが選択した製品群について、前節で算出されたスコアを用いて、その結果をレーダチャートで表示することができる。図 4.16 は生成されたレーダチャートの例である。レーダチャート各軸はユーザによって変更が可能であり、“CPU”、“メモリ”、“デバイス”、“表示機能”、“バッテリー”、“重量”、“価格”の 8 種類がデフォルトである。デフォルトにおける“CPU”、“メモリ”、“デバイス”、“表示機能”については、関連する属性のスコアの平均によって出力される。例えば、“CPU”については“クロック数”と“プロセッサの種類”、“デバイス”については“CD-R/RW”や“DVD-ROM”、“DVD-RAM”などの属性に対するスコアの平均をとる。

表 4.18: 出力形式による読む側のわかりやすさ

形式 (1)	形式 (2)	形式 (3)	形式 (4)
10 点	23 点	18 点	29 点

4.6.4 出力結果に対する評価

次に出力結果に対する評価である．まず要約の出力としてどのような形式がわかりやすいかということをアンケートで調査した．比べたのは以下の 4 種類である．

1. 特徴データを含んだ表のみ
2. 元の製品性能表で抽出された特徴データを強調したもの
3. 特徴データを用いて生成された文章のみ
4. 特徴データを用いた文章と表

被験者は学部および大学院生 8 名である．(1) ~ (4) を順位付けしてもらい，1 位を 4 点，2 位を 3 点，3 位を 2 点，4 位を 1 点としたところ結果は表 4.18 のようになった．このことから文章と表を統合した出力が最もわかりやすいということがいえる．

続いて生成された文章に対する評価を行った．現在文章生成フレームの数は 27 個，説明文は 35 文である．これらは実際の雑誌のレビューなどを基に人手で作成した．また修辞処理用の規則は 8 個である．表 4.13 の特徴データを用いて生成された文章の例を図 4.17 に示す．生成された文章を以下の点について評価した．

評価 (1) 文法的に正しいか

評価 (2) 文章中の接続詞などは正しいか

評価 (3) 文章の流れは正しいか

評価 (4) 冗長なところはないか

評価 (5) 文章として読みやすいか

評価は以下のような 5 段階で行った．

1 点 正しくない

パソコンメーカー各社はそれぞれ、PC1（A社）、PC2（B社）、PC3（C社）を発売した。

基本性能の高いPC3

PC3はCPUにペンティアム III700MHz を搭載している。

PC1もCPUにペンティアム III650MHz を採用しており、比較的高性能だ。

しかし、PC2はHDDに20GBを搭載している。

PC1もHDDに18GBを装備しており、標準以上。

また、PC3、PC1はメモリ（標準）に128MBを搭載。

大容量のメインメモリを装備しているため、使用時の快適度がアップ。

拡張性が高いPC3、PC2

PC2はインターフェースにUSBを3基搭載している。

USBは1つのインターフェイスに最大127台の機器を接続できる。

しかし、PC3はメモリ（最大）が512MBである。

画像処理がよいPC2、PC1

PC2、PC1はインターフェースにIEEE1394を搭載。

IEEE1394は高速でビデオカメラなどとの接続にも向いている。

表示能力のよいPC2

PC2、PC3はグラフィックメモリに8MBを搭載している。

グラフィックメモリが大きければ、高解像度でのフルカラー表示が可能になる。

また、PC2は液晶ディスプレイが15インチTFTである。

携帯性が高いPC3

PC3はサイズが304mm（W）×250mm（D）×33.1mm（H）。

同様に、重さが2.36kg。

この中で最も軽く携帯性に優れている。

PC1も重さが3.1kgと、比較的軽い。

通信機能がよいPC3、PC2

PC2は通信機器に携帯/PHSを採用。

しかし、PC3は通信機器にLANを装備。

LAN端子があるので、ブロードバンドにも対応。

図 4.17: 生成された文章の例

表 4.19: 生成された文章の評価

評価 (1)	評価 (2)	評価 (3)	評価 (4)	評価 (5)
3.7	2.9	3.7	4.3	4.0

2 点 やや正しくない

3 点 正しくない部分もあるが問題はない

4 点 問題はない

5 点 よい

被験者は大学院生 8 名で，評価用のデータとして抽出された特徴データから生成した 3 種類の文章を用いた．結果を表 4.19 に示す．表中の値は被験者の評価の平均をとったものである．評価から考えれば，出力は要約文章としてほぼ問題はないと考えられる．評価 (2) が最も低かった原因は，逆接の接続詞 “しかし” の使用について適切ではない部分があったためであり，修辞処理規則の改良が今後の課題に挙げられるだろう．最終的な HTML 出力を図 4.18 に示す．

4.7 システム全体に対する考察

4.7.1 他種類データへの移行

他種類製品の表を新たに対象とする場合の移行コストについて考察する．ここでは，(1) 表抽出処理，(2) 特徴データ抽出処理，について議論する．

4.7.1.1 表抽出処理の移行

デジタルカメラの製品性能表を対象とし，フィルタリングおよび表領域抽出処理の精度について考察する．キーワード抽出および重み付けのための訓練データとして性能表を含む 50 文書および含まない 50 文書を用意した．また，性能表を含む 93 文書と含まない 100 文書をテストデータとして別に用意した．ここで，PC の場合と大きく異なるのは，性能表を含まない文書には，同一社内で製作されているビデオカメラやスチールカメラの性能表を含んでいることである．

フィルタリングの実験結果を表 4.20 および表 4.21 に示す．表 4.21 のキーワードは， $P(D_{real}|t) > 0.75$ かつ性能表を含む文書群 D_{real} の半分以上の文書に出現した語，ノイズ

パソコンメーカー各社はそれぞれ、PC1 (A社)、PC2 (B社)、PC3 (C社) を発売した。

◆基本性能の高いPC3
 PC3はCPUにペンティアムIII700MHzを搭載している。
 PC1もCPUにペンティアムIII650MHzを採用しており、比較的高性能だ。
 しかし、PC2はHDDに20GBを搭載している。
 PC1もHDDに18GBを装備しており、標準以上。
 また、PC3、PC1はメモリ(標準)に128MBを搭載。
 大容量のメインメモリを装備しているため、使用時の快適度がアップ。

	PC1	PC2	PC3
CPU	650MHz	600MHz	700MHz
メモリ	標準	64MB	128MB
HDD	18GB	20GB	12GB

◆拡張性が高いPC3、PC2
 PC2はインターフェースにUSBを3基搭載している。
 USBは1つのインターフェースに最大127台の機器を接続できる。
 しかし、PC3はメモリ(最大)が512MBである。

	PC1	PC2	PC3
メモリ	最大	256MB	512MB
インタフェース	USB×1, IEEE1394	USB×3, IEEE1394	USB×1

◆画像処理がよいPC2、PC1
 PC2、PC1はインターフェースにIEEE1394を搭載。
 IEEE1394は高速ビデオカメラなどの接続に有利なアイズ

図 4.18: 表と文章を統合した出力

ワードは、 $P(D_{no}|t) > 0.75$ かつ性能表を含まない文書群 D_{no} の $\frac{1}{10}$ 以上に出現した語である。ベイズの定理による手法におけるキーワード t の重みは、 $ws_t^{real} = P(D_{real}|t)$ および $ws_t^{no} = P(D_{no}|t)$ とした。表では、PC を対象とした場合に最も F 値の良かった閾値とデジタルカメラの場合に最も F 値が良かった閾値の結果のみを載せている。

実験結果としては、PC の場合に比べ、精度が若干下がるものの、特にベイズの定理を用いたキーワードと重みでは、高い精度を得ることができていることが確認された。性能表を含まない文書群には、ビデオカメラやスチールカメラのような記述内容が似通った性能表が含まれているにも関わらず、適切なフィルタリングが行われていることが実証された。エントロピーを用いた手法とベイズの定理を用いた手法の精度差の原因は、PC の場合と同じく、重みが正規化されているか否かによるものだと考えられる。

続いて、表領域抽出について述べる。ここでも、フィルタリングによって得られた結果で最も精度が良いものを用いた。それぞれの実験結果を表 4.22 および表 4.23 に示す。表領域抽出においては、フィルタリングに比べ、再現率・適合率とも大きく下がった。これ

表 4.20: エントロピー $M_{real} = 30, M_{no} = 15$ - デジタルカメラ -

閾値 $th1$	再現率	適合率	F 値
0.25	100.0%	65.0%	82.3%
0.80	94.6%	75.9%	86.1%

表 4.21: ベイズの定理 $M_{real} = 29, M_{no} = 23$ - デジタルカメラ -

閾値 $th1$	再現率	適合率	F 値
0.50	96.8%	92.8%	95.1%
0.55	96.8%	94.7%	96.0%

も、PC の場合と同じく、表領域抽出のための閾値の設定に問題があるためだと考えられる。

総じて、PC の場合と比べて、若干の精度差があるが、高い再現率および適合率を得ることができた。問題点としては、製品ごとにフィルタリングおよび表領域抽出で用いる閾値の値が変動することにある。この点においては、検証した 2 つの手法以外の重み付けアルゴリズムによって、さらなる実験および考察が必要である。

4.7.1.2 特徴データ抽出処理の移行

次に、特徴データ抽出処理の移行コストについて考察する。ここでは、携帯電話とデジタルカメラの製品性能表を新たに対象とした。

1. 携帯電話に関する製品性能表

携帯電話に関する製品性能表から人手によって単位の分類とキーワードの抽出を行った。単位の分類はパソコンに関する製品性能表からの特徴抽出処理と同様、比較処理のための分類（表 4.9）と標準値算出のための分類（表 4.10）の 2 種類を行った。それぞれを表 4.24 および表 4.25 に示す。文字データの比較処理で用いられるキーワードの例としては、“画像”、“反射型”などが挙げられる。現在登録されてあるキーワードは 23 個である。関連項目はない。

2. デジタルカメラに関する製品性能表

携帯電話の場合と同様に人手で単位の分類とキーワードの抽出を行った。単位の分

表 4.22: 表領域抽出処理 - エントロピー・デジタルカメラ -

閾値 $th1$	再現率	適合率
0.80	80.7%	86.6%

表 4.23: 表領域抽出処理 - ベイズの定理・デジタルカメラ -

閾値 $th1$	再現率	適合率
0.55	82.8%	95.1%

類は携帯電話と同様に比較のための分類と標準値算出のための分類を行った。それぞれを表 4.26 および表 4.27 に示す。標準値の算出方法はパソコンおよび携帯電話の場合と同じく平均値と最頻度によって求められるが、比較処理のための分類は新たな比較方法が追加された。既存の「大きいほどよい単位」, 「小さいほどよい単位」, 「属性によって変わる単位」の他に新たに追加されたのは、「範囲が広いほどよい単位」と「データ数が多いほどよい単位」の 2 種類である。また、パソコンおよび携帯電話の場合ではなかったが、単位“mm”のように、一つの単位が複数の比較方法に分類される場合もある。文字データの比較処理で用いられるキーワードの例としては、“動画”, “内蔵メモリ”などが挙げられる。現在登録されてあるキーワードは 27 個である。携帯電話の場合と同じく関連項目はない。

携帯電話に関する製品性能表を用いた特徴抽出処理は上記の辞書を書き換えるのみで移行できた。デジタルカメラを対象とした場合は、パソコンおよび携帯電話の場合と比べ、比較処理のための分類が増えたが、移行処理は文書を対象とした従来の情報抽出処理で用いるテンプレートの作成などの作業と比べ、遙かに容易に行えることが確認された。

携帯電話については 12 機種 of 製品性能表を 4 つに分け、デジタルカメラについては 17 機種を 5 つに分けて特徴データ抽出処理について実験を行った。雑誌のレビューを基準に評価した結果、パソコンの製品性能表を対象とした場合と同様に妥当な特徴データが抽出されていることを確認した。

表 4.24: 比較のための単位の分類 - 携帯電話 -

数値が大きいほどよい単位	階調, 色, 件, 時間, 和音など
数値が小さいほどよい単位	円, g
属性によって変わる単位	mm

表 4.25: 標準値の算出処理のための分類 - 携帯電話 -

平均値	件, 円, mm, 時間, g, 分など
最頻度	階調, 色, ドット, 和音など

表 4.26: 比較のための単位の分類 - デジタルカメラ -

数値が大きいほどよい単位	画素, 倍, ピクセル, 枚など
数値が小さいほどよい単位	g, F, 円 など
属性によって変わる単位	mm, m, 時間, 秒など
範囲が広いほどよい単位	F, 秒, mm, m など
データ数が多いほどよい単位	ピクセル, F など

表 4.27: 標準値の算出処理のための分類 - デジタルカメラ -

平均値	円, cm, mm, 時間, 画素, 倍など
最頻度	インチ, ドット, ピクセル, 分など

4.7.2 関連研究との比較

次に、システム全体について、他研究と比較・考察する。Web上から製品に関する情報を抽出し、それらを用いて製品購入の支援をするという観点から考えると、本研究はオンラインショッピングに関するシステムと比較することができる。関連研究の節でも述べたが、オンラインショッピングに関する研究としては、Doorenbosら[9]によるShopBotが有名である。ShopBotは、Web上の複数のオンラインショップで、商品の値段を比較して、最も安い商品を見つける買い物エージェントである。与えられたオンラインショップのサイトからどのようにして、目的の商品の価格を抽出するかを学習する。同様に、複数のオークションサイトからユーザの希望する製品を、ユーザの希望する落札価格の範囲内で、なるべく安価で落札するBiddingBotなども提案されている[23]。これらの研究は、主に価格情報の抽出に重点を置いている。これらの手法は、我々の提案した製品選択支援システムにも製品の価格抽出という点で有用であるが、代替にはならない。

一方で、Chaiらの提案した、対話的にユーザの求める製品を検索するシステムHappyAssistantがある[4]。このシステムとの相違点を以下に述べる。

(1) HappyAssistantは同一社内の製品のみを扱う。

これに対し、我々のシステムは、複数社の製品を比較し、それぞれの相対的な特徴を提示することができる。また、HappyAssistantでは、いかにWeb上から製品の情報を抽出するかについては議論されていない。

(2) HappyAssistantは得られた結果に対して、簡単な説明文のみを出力する。

一方、我々のシステムは、製品の相対的な特徴を文章で要約し、表やグラフも提示することができ、よりわかりやすい情報提示が可能である。

(3) HappyAssistantは対話によって製品検索ができる。

我々のシステムは、5つのあらかじめ定義された要求とユーザによる重み選択、適合性フィードバックによってある程度柔軟な製品検索が可能である。しかし、対話を用いれば、理論的には、より柔軟な検索が可能である。この点については次節で述べる。

4.7.3 対話システムと文脈処理

対話によって製品検索を行うことの有効性は先行研究より明らかである [4]。製品選択支援システムは現在、ユーザからの要求と適合性フィードバックによって柔軟でかつ有効な製品検索が行えるが、より柔軟さを求めれば、システムとの対話が欠かせない。

本論文では、本章の製品選択支援システムとは対象が異なるが、文脈情報を参照しながら、小学校1年生の算数のドリルテキストを対話処理によって解くシステムを3章で述べた。この対話支援型問題解決システムの目的は、3.5.4節でも述べたように、算数のドリルテキストを解くシステムを構築することではなく、ある問題を対話を用いて解決することにある。すなわち、このシステムで用いた対話型の問題解決手法は、製品選択支援システムを対話型のシステムへ拡張する際にも有効である。本節では、3章で述べたの考え方をふまえ、製品選択支援システムを拡張した場合の対話処理の例を示す。

3章で述べたシステムは、あらかじめ定義されたコマンドを組み合わせることによって、問題解決を図る(3.2.1節および3.3.3節を参照)。製品選択支援システムで考えると、これはユーザの要求に置き換えることができる。このユーザの要求をHappyAssistantに倣い、コンセプトと呼ぶ。対話に応じてこのコンセプトを組み合わせ、製品検索を行う。コンセプトは、コンセプト名と関連する表の属性で構成できると考えられる。すなわち、これは表4.11に該当する。既存のシステムでは、コンセプトは一度しか入力できなかったが、対話を用いれば、その内容に応じて複数のコンセプトを満たす検索が可能になる。

続いて、文脈表現について考える。3.3.4節で示したように、3章で述べたシステムでは、文脈は先行対話文に関する情報、注意状態(焦点要素)およびその時点までに得られたコマンドのリストで構成される。製品選択支援システムを対話システムに拡張する際には、注意状態およびコンセプト(3章ではコマンド)のリストが重要になる。

ここで、「ノートPC」や「高性能」、「携帯性」などの幾つかのコンセプトがあらかじめ定義されていると仮定し、図4.19のような対話例を考えよう。ここで、Sはシステムの発話、Uはユーザの発話を表す。U1の解析結果により、コンセプトとして「ノートPC」が抽出される。続いて、U2でコンセプト「高性能」が抽出される。システムから他に制限がないかの確認が行われ、U3でより具体的な制約がかかる。ここで、「20万円以下」や「DVD-RAM」といったコンセプトより具体的な要求が抽出される。これは、文脈的には焦点要素だと考えることができる。この結果を踏まえ、コンセプト「ノートPC」「高性能」を満たし、焦点要素「20万円以下」や「DVD-RAM」も同時に満たすような、製品が検索され、4.6節で述べたような流れによって、結果が出力される。U4では、文脈処理によって「2番目」が「PC2」であると解釈され、「PC2」が焦点要素とし

S1: ノートPC とデスクトップPC のどちらを探していますか?
U1: ノートで。
S2: 何か希望がありますか?
U2: 高性能なものがいいです。
S3: 他に制限などはありますか?
U3: 20万円ぐらいまでで、DVD-RAM がついていて欲しいかな。
S4: 候補としてPC1 とPC2, PC3 があります。
PC1 は.... PC2 は.... PC3 は.... (要約文などの出力)
U4: 2番目みたいなやつで、もう少し軽いのはありますか?
S5: これはどうでしょう。

図 4.19: 製品選択支援システムにおける対話例

てセットされる。さらにコンセプト「携帯性」が得られる。U3 までに得られたコンセプト群や発話内容と U4 で新たに得られたコンセプトおよび焦点要素を考慮した検索を行うことで、新たな検索結果をユーザに提示する。

このように、3章で述べた考え方を参考にして、製品選択支援システムを対話型のシステムへ拡張することが可能である。実際には、対話のプランニングや戦略などについてさらに議論する必要があるだろう。

4.8 おわりに

本章では、Web 上の製品性能表を用いた情報抽出・検索・統合・要約手法について述べた。有効性を検証するために作成したプロトタイプシステムはユーザの意図を考慮した製品選択が行え、文章やグラフへの要約処理により、ユーザにとってよりわかりやすい出力を提示することができる。

Web からの表抽出および表解析については、高い精度が得られた。表からの特徴抽出処理についても十分な再現率と適合率が得られている。製品選択支援における検索処理においても、5つの要求と属性の重みを自由に変えられることで、柔軟な検索が行え、適合性フィードバックによって再現率および適合率の大幅な上昇が確認された。表による出力だけでなく、特徴抽出・スコアリングの結果を用いて、文章・表・グラフの様々な形式で出力することが可能であり、よりわかりやすい情報提供が可能である。また、表を扱うという利点から、従来のテキストベースの情報抽出システムと比較して、数少ないルールや

パターン，知識で実現でき，実際に他種類製品に本手法を適用することで，その有効性が確認された．

今後の課題としては，

- (1) 知識の自動構築
- (2) 対話による製品検索
- (3) 画像との統合

の 3 点が挙げられる．

Web 上からの性能表抽出および特徴データ抽出処理については，(1) について考察したが，表構造生成処理で用いられる正規化のための辞書や文章生成で用いられる文章生成フレームの獲得などについては，本論文で議論していない．表構造生成処理で用いる辞書については，正規化されていない表構造の要素をベクトル化し，ベクトル空間で類似度を測定することによって，表記の揺れや同義語の獲得を行っている．文章生成フレームの追加・修正作業については，製品を紹介した Web サイトから文章生成フレームになりうる候補を自動的に抽出し，それを基に GUI を通してユーザとのインタラクションをとりながら，効率的に文章生成フレームを獲得するツールの開発を進めている．(2) については，4.7.3 節で具体例を述べた．対話のプランニングなどを含めて，現在システムを拡張中である．(3) については，統合のための画像抽出および分類手法について，次章で詳しく述べる．

第 5 章

文章および画像からの情報抽出

Web 上の画像とその画像に隣接する文章には関係があることが多い。この関係をメディア間の文脈と捉え、文章間で先行詞と照応詞の関係を特定するように、隣接する文章から抽出された情報を用いて画像の内容を特定する手法について述べる。

5.1 はじめに

4章で述べたように、インターネットの普及により多くの文書が電子化されている。電子化されている文書が文章や表のみで構成されていることは稀で、多くの場合、文章や表とともに画像が用いられる。メディア間の文脈を考えた場合、画像内容の特定は重要な課題となる。また、4章で述べたように、我々は表とその表から生成された文章、グラフを用いて複数の出力形式を統合した要約処理の実現を目指している。現在の表や文章グラフによる要約出力に、関連する画像が追加されれば、ユーザにとってよりわかりやすい要約を提示できることは4章の実験結果より明らかである。本章では、複数の出力形式による要約処理への統合を考慮した Web 上の製品紹介サイト中の画像の内容推定手法について提案する。

画像内容の特定については、画像検索を目的とした多くの研究がなされている [27][31][32][64]。これらの手法の多くは、画像のエッジやテクスチャ、色情報などによってその画像の特徴を抽出している。しかし、Web 上の画像は、解像度が十分であるとはいえず、従来の画像解析手法は必ずしも適さない。

一方で、Web 上の画像とその画像に隣接する文章には関係があることが多い。実際に検索エンジン Google などでは、Web 上の文章などから解析されたデータを基に、画像検索を行うシステムが公開されている [17]。隣接する文章を用いることで、Web 上の低解像度な画像に対しても、従来の画像解析手法を用いずにその画像内容を特定すること

が可能となる．文章と画像間の関係を用いた画像内容の特定に関する研究も多く存在する [15][20][62][63][65][74]-[76][79]．Houghton ら [20] や Srihari[63]，山田ら [79] の目的は新聞記事などの文章と写真中の人物の特定であり，製品紹介のサイト中に存在する画像の内容を特定しようとする我々の目的とは異なる．Favela ら [15] や Swain ら [65]，渡辺ら [74]-[76] の目的は画像と文章との統合や検索システムの構築などであるが，画像中に出現するテキストや色情報などを用いている．我々が提案する手法は画像サイズを抽出段階で用いるが，それ以外の画像情報を内容特定処理では用いない．

我々の目的は先にも述べたように，4章で構築した表や文章，グラフなどの複数の形式による要約出力に画像を統合し，マルチメディア要約処理を実現することである．そこで，製品紹介サイト中の画像を，接続する文章を用いて4章で説明した文章生成処理の「話題」のどれに関連するかを分類する手法を提案する．具体的には，(1) 人手によるキーワード，(2) $tf \cdot idf$ 法によって重み付けされたキーワード，(3) ベイズの定理によって重み付けされたキーワード，を用いた分類手法の精度を検証する．以下，5.2節では，Web サイトからの画像および隣接文章の抽出手法について，5.3節ではキーワード抽出とその重み付けについて，5.4節ではそれらキーワードを用いた評価実験について述べ，5.5節でまとめる．

5.2 画像と隣接文章の抽出

我々が対象とするのは図 5.1のような製品の特徴や装備を紹介する企業の Web ページである．本節では，このような Web ページの HTML から分類処理の対象となる画像とそれらに隣接する文章を抽出する手法について述べる．

5.2.1 対象画像の抽出

Web ページ中の画像は大きく二つに分けられる．図 5.2はそれぞれのタイプに属する画像の例を示す．

タイプ 1 文字やボタンなどの画像

タイプ 2 製品に関連する画像

一般に Web ページでは，見栄えをよくするために，見出し文字やリンクなどに画像を使うことが多い．タイプ 1 の画像はそれらに属し，本研究が対象とする画像とは異なる．故に Web ページ中の画像がタイプ 1 の画像か，タイプ 2 の画像かを判別する必要がある．判別処理には画像のサイズを用いる．判別条件は以下の 2 つである．

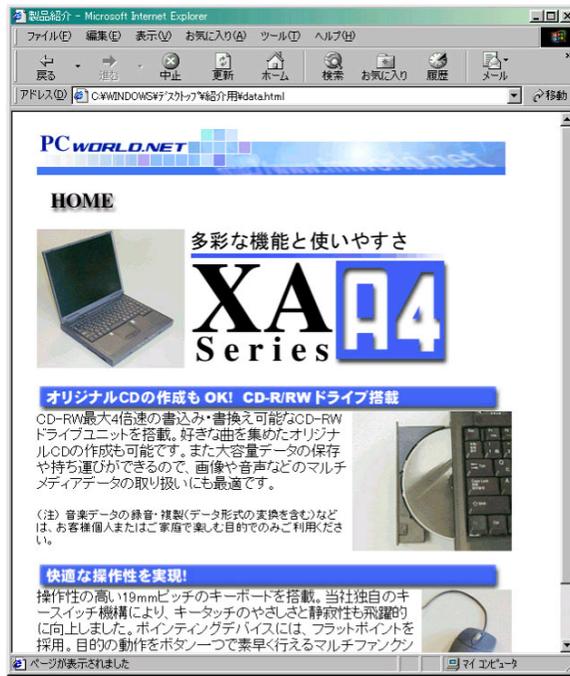


図 5.1: 製品紹介サイト

1. 画像のサイズが 55 × 55 ピクセル以上の画像
2. 画像の縦横比が 3.8 以下の画像

上記の 2 つの条件を満たす画像をタイプ 2 の画像として抽出する．これらの条件は 337 画像から人手により作成された．また，Web ではサイトの製作者がネットワークの通信速度などを考慮して，1 枚の大きな画像を複数に分割して保存することがある．そこで本システムは，密着する画像同士は 1 枚の画像として扱う．



図 5.2: 画像のタイプ

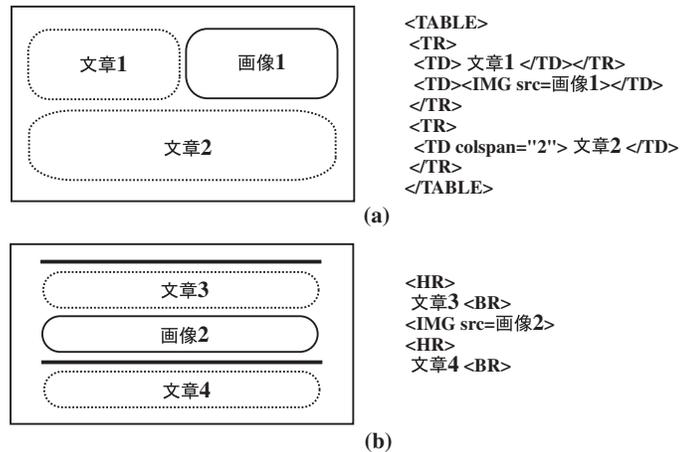


図 5.3: レイアウトとその HTML 表記

5.2.2 隣接する文章の抽出

続いて，抽出した画像に隣接する文章を抽出する．隣接する文章の抽出処理の流れを以下に示す．

1. HTML のソースから対象となる画像ファイルの タグを特定する．
2. その 周辺の文章を抽出する．
3. 中に ALT データが存在すれば，その ALT データの内容を含む隣接テキストを対象文章として抽出する．
4. タグからレイアウト情報を抽出する．
5. 抽出されたデータから最も画像に近く，優先順位の高いタグでレイアウトされたテキストを対象文章として抽出する．

現在，レイアウト解析で考慮しているタグは<TABLE>，
，<HR> などである．タグの優先度は以下の通りである．

<TABLE> >
 > <HR>

図 5.3は実際のレイアウトとその HTML 表記の例を示す．図 5.3において，画像 1 および画像 2 がそれぞれ対象画像だと仮定すると，文章 1 および文章 3 がそれぞれの画像の隣接文章となる．

表 5.1: タグの重み

タグ	n
フォント色変更 	1.5
強調 	1.5
フォント拡大 	1.75
フォント縮小 	0.75
フォント拡大 and (色変更 or 強調)	2.0

5.3 キーワードと重みづけ

画像分類のためのキーワードを定義する．我々の目的は Web 上の画像が表 4.17 に示された話題のどれに分類されるかを判定することにある．キーワードと重みは (1) 人手によるキーワード，(2) $tf \cdot idf$ 法によって重み付けされたキーワード，(3) ベイズの定理によって重み付けされたキーワードの 3 種類がある．

5.3.1 人手によるキーワードと重み

まず，人手によるキーワードと重み付けについて述べる．キーワードの例を以下に示す．
クラス [画像処理]：

(画像 | グラフィック)(を)?(扱う | あつかう | 処理 | 処理する)? ボード

ここで，記号 “?” はその単語が 1 回もしくは 0 回生じることを示し，“|” は or を示す．分類処理はこれらの定義されたキーワードと隣接文章の第 1 文目とのマッチングを行う．キーワードが文章中の「を格」とマッチングした場合，そのキーワードが分類されている項目を画像のクラスとして決定する．また，テンプレートの「を格」とマッチングしない場合でも，第 1 文でマッチしたキーワードはそれ以降の文でマッチングした場合の 2 倍の得点を加算する．また，フォントサイズの変更や文字の強調などのタグに囲まれたキーワードがあった場合は，そのタグに合わせて得点が n 倍される．タグごとの n の値の一覧を表 5.1 に示す．第 1 文とのキーワードマッチング終了後，それ以降の文章もキーワードのマッチングを行う．タグが存在すれば，同様に得点を n 倍する．全ての文章とマッチングを終えた後，総得点の最も多い話題をその画像のクラスとする．総得点と同値の場合，第 1 文でマッチしたキーワードの多い方をその画像のクラスとする．

5.3.2 $tf \cdot idf$ 法による重み

$tf \cdot idf$ 法は一般によく用いられる重み付け法である． $tf \cdot idf$ は 1 文書中に存在する単語の頻度とその単語の出現する文書数によって求められる．ある文書 d 中に存在するキーワード k の頻度 (term frequency) を tf_k と表す．一方で．どの文書にも高頻度で出現する語は網羅性は高いが各文書の特徴づける語としては役に立たない．特定性表す尺度としては idf (inverse document frequency) が知られている．

$$idf_k = \log \frac{N}{df_k}$$

ここで， N は全文書数， df_k はキーワード k が出現する文書数である． $tf \cdot idf$ 法はこの二つを掛けあわせることによって実現される．

$$w_k = tf_k \cdot idf_k$$

$tf \cdot idf$ 法によって重み付けされるキーワードは以下の手順で各文章から抽出される．

1. 字種ごとに分割する．
2. ひらがなを削除する．
3. 全角記号を削除する．
4. 1文字のみで構成される語を削除する．
(例) 高い → 高
5. 空白や記号で分割する．
(例) CD-R/RW → CD-R, RW
6. 数字を含むものを削除する．
但し，数字を含んでも削除されない語も定義されている．
(例) IEEE1394 や MP3 など
7. 不要語リストに含まれている語は削除する
(不要語) for, I, my など

図 5.4にキーワード抽出処理の例を示す．この手順によって抽出されたキーワードは後述するベイズの定理を用いた重み付け処理でも用いられる．

続いて，実際の $tf \cdot idf$ 法による重み付けの手順について述べる．ここで文書とは画像に隣接する文章群のことを指す．

Sentence: DVD-R/RW も読み込める 2 倍速の DVD-R を搭載
 Step1 : DVD-R/RW も読み込める 2 倍速の DVD-R を搭載
 Step2 : DVD-R/RW 読込 2 倍速 DVD-R 搭載
 Step3 : DVD-R/RW 読込 2 倍速 DVD-R 搭載
 Step4 : DVD-R/RW 2 倍速 DVD-R 搭載
 Step5 : DVD-R RW 2 倍速 DVD-R 搭載
 Step6 : DVD-R RW 倍速 DVD-R 搭載

図 5.4: キーワード抽出処理の例

1. 各々の文書におけるキーワードの重みを計算

$$dw(d_i, k) = w \cdot tf_k \cdot \log \frac{N}{df_k}$$

ここで, $dw(d_i, k)$ は文書 i におけるキーワード k の重み, w は文書中の第 1 文目の重み, tf_k はキーワード k の出現頻度, df_k はキーワード k を含む文書数, N は文書の総数を表す.

2. 分類項目ごとの重みを計算

$$cw(C_j, k) = \sum_{k \in C_j} dw(d_i, k) \cdot \frac{df(C_j, k)}{df_k}$$

ここで, $cw(C_j, k)$ はクラス C_j におけるキーワード k の重み, $df(C_j, k)$ はクラス C_j 中でキーワード k を含む文書の総数を表す.

$cw(C_j, k)$ が $tf \cdot idf$ 法による各クラスごとの最終的なキーワードの重みとなる. 分類処理では文書中のキーワードを抽出し, 各クラスごとの重みの総和を求め, 和が最大となるクラスをその画像のクラスとする.

5.3.3 ベイズの定理による重み

別の重み付け手法として, 前章の性能表抽出で用いたベイズの定理を採用した. 性能表抽出処理の場合と同様に, $C = [C_j]_{j=1}^M$ において, 事前確率 $P(C_j)$ と条件付き確率密度分布 $p(k|C_j)$ から, あるキーワード k がクラス C_j に属する事後確率 $P(C_j|k)$ は次の式で求められる.

$$P(C_j|k) = \frac{P(C_j)p(k|C_j)}{p(k)}$$

表 5.2: 対象画像抽出処理の結果

	訓練データ	テストデータ
再現率	98.0%	97.2%
適合率	91.5%	94.0%

我々は上記の式で得られた事後確率 $P(C_j|k)$ をキーワード k の重みとする．分類処理では， $tf \cdot idf$ 法と同様に，文書中のキーワードを抽出し，各クラスごとの重みの総和を求め，和が最大となるクラスをその画像のクラスとする．

5.4 評価実験

本節では，対象画像の抽出処理と 3 種類のキーワードと重みを用いた画像分類処理の精度比較・考察を行う．

5.4.1 画像抽出処理

画像抽出処理のための抽出条件作成には，訓練データとして 337 画像を用いた．条件作成に用いていないテストデータは 450 画像である．表 5.2 に作成された抽出条件による訓練データおよびテストデータの実験結果を示す．評価基準は前章でも用いた適合率と再現率である．

訓練データ・テストデータ共に，高い適合率と再現率が得られた．画像抽出処理は，分類処理の前処理であるため，高い再現率が望ましい．作成された抽出条件はシンプルだが，十分な精度が得られ，本手法の有効性が確認された．また，作成された条件は画像のサイズのみを用いており，対象とする製品の種類等に依存しない．そのため，本条件を用いた画像抽出は，他種類の製品紹介サイトに適用することができる．

5.4.2 画像分類処理

抽出処理で使用した 787 画像の内，隣接する文章がある 325 画像を分類処理実験に使用した．ここで，「文章がある」とは 2 単語以上で構成される文字列のことを指す．図 5.5 は画像とそれに隣接する文章，分類されるべきクラス名の例である．人手によるキーワード作成， $tf \cdot idf$ 法およびベイズの定理を用いた重み付け処理のために用いられた訓練データの数は 182 文書である．人手によるキーワードは 184 語， $tf \cdot idf$ 法およびベ



図 5.5: 画像とそれに隣接する文章

ズの定理による重み付け処理で用いたキーワードは 1111 語であった．実験結果を表 5.3 に示す．表中の w は第 1 文目の重みを表している． $tf \cdot idf$ 法については，訓練データ中の各クラスに属する文書の頻度が大きく異なるため，以下の重みの 2 式で正規化を行った．

$$\text{正規化1} = 1 - \frac{df_c}{N} \quad \text{正規化2} = \log \frac{N}{df_c}$$

ここで， N は全文書数， df_c は各クラスに属する文書数である．結果としてベイズの定理を用いた重み付け手法が最も良い精度となった． $tf \cdot idf$ 法についても比較的高い精度を得られており，人手によるキーワードよりも有効であることがわかった．正規化についても一定の効果が認められた． $tf \cdot idf$ 法による重み付けの精度がベイズの定理によるものよりも低い原因としては，ベイズの定理による重みづけでは，その重みは確率を基にしているため値の範囲が 0 から 1 までであるのに対して， $tf \cdot idf$ 法は重みが理論的には 0 から無限大までの値をとるため，過学習が起きたためだと考えられる．ベイズの定理による重みが第 1 文目の重みを大きくしても精度にあまり差がでないのに対して， $tf \cdot idf$ 法による重みはあまり極端に大きな値にすると精度が落ちる点，重みの正規化を行うと精度が上がる点も，これが原因だと考えられる．

続いて，テストデータを用いてそれぞれの手法の比較評価を行った．実験結果を表 5.4 に示す．テストデータに対しても， $tf \cdot idf$ およびベイズの定理による重み付け手法の方が人手によって作成されたキーワードと重みよりも有効であることがわかった． $tf \cdot idf$ 法とベイズの定理による重み付けの優劣をみると，訓練データの場合と逆の結果となった．これは，学習されたキーワードがテキスト中に多く存在しない場合，訓練データとは逆に

表 5.3: 画像分類処理の結果 - 訓練データ -

		$w = 1$	$w = 2$	$w = 5$	$w = 10$	$w = 20$
$tf \cdot idf$ 法	正規化なし	89.0%	90.1%	91.2%	90.1%	87.9%
	正規化 1	90.1%	90.7%	91.8%	89.5%	90.1%
	正規化 2	90.7%	92.9%	92.3%	91.8%	90.1%
ベイズの定理		97.3%	98.9%	99.5%	99.5%	99.5%
人手によるキーワード		78.0%				

表 5.4: 画像分類処理の結果 - テストデータ -

		$w = 1$	$w = 2$	$w = 5$	$w = 10$	$w = 20$
$tf \cdot idf$ 法	正規化なし	70.6%	72.0%	68.5%	68.5%	66.4%
	正規化 1	70.6%	73.4%	70.6%	69.9%	67.8%
	正規化 2	78.3%	79.0%	76.2%	73.4%	70.6%
ベイズの定理		74.1%	74.8%	75.5%	74.8%	72.7%
人手によるキーワード		66.4%				

過学習された大きな重みを持つキーワードが有効に作用したためだと考えられる。また、正規化に関しては、訓練データの場合に比べ、明らかに有効に働いていることがわかる。テストデータを用いて重みを再学習したところ、 $tf \cdot idf$ 法による重み付け（正規化 2, $w = 2$ ）では、89.5%、ベイズの定理による重み付け（ $w = 5$ ）では、97.2% となり、訓練データの場合同様にベイズの定理による重み付けの方が高い精度を収めた。

また、キーワードの数についても考察する。 $tf \cdot idf$ 値を基準にキーワードを削除した。その結果を図 5.6 に示す。実験結果からわかるように、有効なキーワード削減が行えているとはいえない。特にテストデータに関しては、十分な精度が得られなかった。

総括すると、 $tf \cdot idf$ 法およびベイズの定理による重み付け法は妥当な重み付けを行ったことが言える。また、これらの手法が訓練データおよびテストデータの双方で人手による重みに比べて有効であることが確認された。ベイズの定理による重み付けに関しては十分な学習データが存在した場合は最も有効であることがわかった。一方で、キーワードが十分でないケースでは、重みの値が大きくなる $tf \cdot idf$ 法の方が有効であることが確認された。しかし、キーワードが十分に存在した場合のベイズの定理による重み付けの精度が

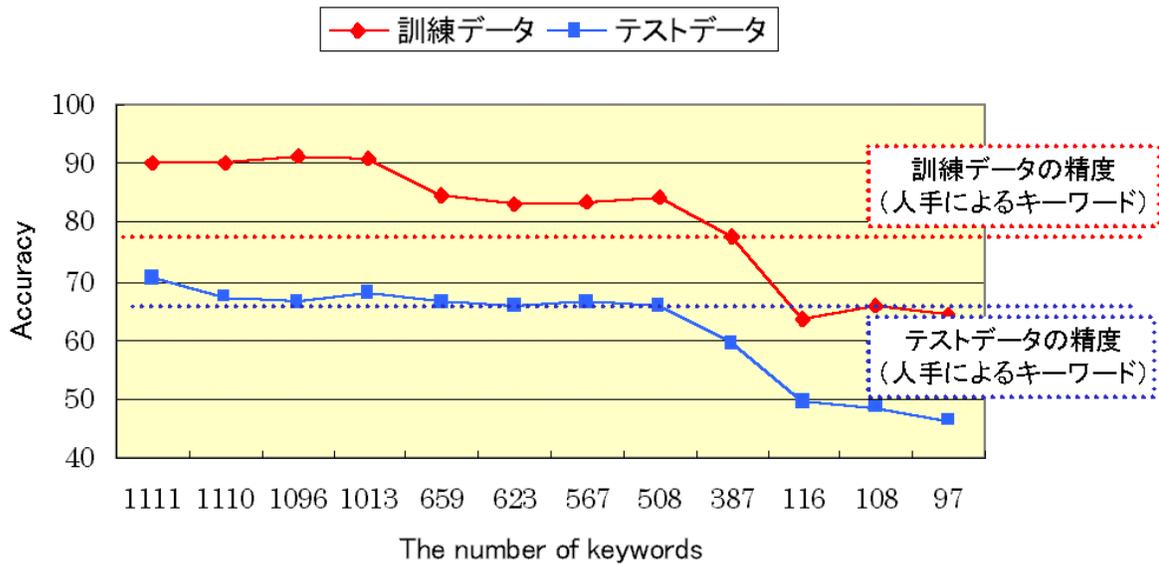


図 5.6: キーワード数と精度

極端に高いため、キーワードの追加が今後の大きな課題となる。また、キーワードの追加に伴う計算量の増加から、必要なキーワードと不必要なキーワードとの判別も今後の重要な課題である。

5.5 おわりに

本章では、製品紹介サイト中の画像の内容を隣接する文章を用いて分類する手法について提案した。画像自体を解析せず、単純なパターンマッチのみで比較的高い精度を得ることができ、本手法の有効性が確認された。3種類の重みづけ手法を用いて比較実験を行ったところ、人手によるキーワードと重みよりも $tf \cdot idf$ 法およびベイズの定理による重みづけ法の方が有効であることが確認された。今後の課題としては、分類精度向上のための有効なキーワードの抽出、前章で述べた表の特徴抽出・要約処理と本手法を統合したマルチメディア要約処理への拡張などが挙げられる。

第 6 章

結論

本論文では、対話からの発話意図や焦点要素などの情報抽出，Web 上からの表の抽出と構造解析，情報統合，形式変換による要約処理，文章や画像からの情報抽出手法について述べた．それらの抽出された情報を用いた文脈処理についても議論した．文章間などのメディア内の脈略だけでなく，表と文章や文章と画像など異なる表現形式間やメディア間での脈略も文脈として考え，複数の情報を統合して解釈および生成することの有効性を明らかにした．

2章では，本論文で述べる対話支援型問題解決システムおよび製品選択支援システムと，それらのシステムで扱われる文脈情報の基本的な考え方について述べた．

3章では，問題解決過程における対話からの情報抽出および抽出された結果を用いた文脈処理について述べた．対話文の解釈・生成手法と対話における文脈表現の定義を行った．定義された文脈表現を用い，小学校 1 年生の算数のドリルテキストを対象とした対話処理システムを開発し，そのシステムについて評価した．システムは問題解決のための処理手続きをコマンドフレームで表現し，カレントコマンドの付加手続きに不具合が生じると，このフレームの格枠を利用して発話文を生成し，教師からの発話を求める．本システムでは対象の性質上，システムを生徒，人間が教師という役割で構築されているが，本研究の目的は，生徒の役割を果たすシステムを構築することではなく，自然言語による対話処理機能を備えた問題解決システムの構築にある．定義された文脈表現は他の対話処理にも利用可能であると考えられる．またシステム自体も，使用しているコマンドフレームを定義し直すことで他分野の問題解決にも適用可能であると思われる．本論文では，自然言語による対話を対象としているが，実際の問題解決においては，音声やジェスチャも併用される．入力もしくは出力をマルチモーダル化するために，定義された文脈表現にさらなる考察が必要になるだろう．

4章では，表を主な処理対象とする情報抽出手法，表を文章といった他形式に変換する

手法などを提案し，その応用例として，表から抽出した情報を用いた製品選択支援システムについて述べた．文章ではなく表を情報抽出の対象として考える利点としては，

- 表は既に要約された文章に比べ，多くの情報を含んでいる．
- 文章からの情報抽出と異なり，任意の製品群から様々な特徴データを抽出することができる．
- 文章からの情報抽出システムに比べ，少ないパターンやルールで実現ができ，他種類の製品にシステムを移行する際のコストが小さい．

などが挙げられる．

プロトタイプシステムにおいて，表の抽出および解析，特徴データ抽出処理，ユーザの意図を考慮した検索，出力形式のわかりやすさ，他製品への移行コストなどについて評価・考察を行った．表の抽出および解析処理では，高い精度を得た．特に表の構造を解析し，2次元の表を1次元のリスト形式に変換する表構造生成処理では，文書中のHTMLに文法ミスが存在しなければ，100%の変換率に達することが確認された．特徴データ抽出処理では，雑誌による評価と被験者による評価の双方において，本手法の有効性が確認された．ユーザの意図を考慮した製品検索処理においても，5つの要求と属性の重みを自由に変えられることで，柔軟な検索が行え，適合性フィードバックによって再現率および適合率の大幅な上昇が確認された．出力形式については，表単体，文章単体よりも表や文章などを統合した出力結果の方がユーザにとってわかりやすいことが示され，他形式への変換，複数の出力形式を統合することの重要性が確認された．また，生成された文章についても要約として機能することが実験により確認された．他製品への移行コストについては，表抽出に必要なキーワードと重み付け処理や対象を携帯電話およびデジタルカメラとした場合の特徴データ抽出処理における変更点を調査し，幾つかの修正・追加が必要であるが，容易に移行できることが確認され，本手法の有効性が確認された．

現在のシステムは，要求と重みの選択，結果の適合性判断によって製品選択を支援しているが，より柔軟でユーザフレンドリな検索を行うことを考えれば，対話による検索方式の構築が必要となる．この対話による検索方式を実現するためには，対話の文脈の利用，プランニングなどが不可欠である．これら文脈処理には3章で提案した文脈表現を使用することが可能であると考え，その具体例を示した．

また本システムは現在，結果の出力の際，表，文章，グラフの形式で出力できるが，文章や表だけではユーザに伝わらないもしくは伝わりにくい情報も存在する．例えば，外観に関する情報などである．よりわかりやすい出力形式を実現するためには，画像との統合が不可欠である．そのための画像内容の推定手法については5章で述べた．

5章では、画像に隣接する文章をレイアウト情報などを用いて抽出し、それらを用いた画像内容の分類手法について提案し、実験・評価を行った。対象としている Web 上の画像は通信速度や表示などの問題から、従来の画像処理手法に用いるために十分な解像度があるとはいえない。一方で、Web 上に存在する表や文章、画像などの各メディア間には関連性がある。そこで、隣接する文章を用いて画像内容を推定することは、有効であるといえる。また、実際に従来の画像解析手法を用いるよりも低コストで実現ができる。

どのようなクラスに分類するかについては幾つか考えがあるが、ここでは4章の製品選択支援システムにおける出力形式との関連で、4章で生成される文章の「話題」をクラスとした。そのクラスに画像を分類するための手法として、(1) 人手によるキーワード、(2) $tf \cdot idf$ 法、(3) ベイズの定理、の3種類のキーワードセットを用意し、比較・考察した。実験結果より、作成するためのコスト、精度の両面とも、人手によるキーワードよりも $tf \cdot idf$ 法およびベイズの定理を利用したキーワードと重みが優った。訓練データに関してはベイズの定理を利用したキーワード、テストデータに関しては $tf \cdot idf$ 法を利用したキーワードの精度が最も良くなり、訓練データの数を増やせば、ベイズの定理を用いた手法が最も有効であることが確認された。

訓練データを増やした場合、キーワードの数も増加するという問題が生じる。キーワードの数の削減についても実験し、考察したが、有効な結果が得られなかった。この点については今後の課題である。また、5章で述べた画像分類は、4章の要約処理と関連がある。分類された画像と4章で生成された文章との統合手法についても今後の課題となる。統合の際には今回の分類とは別に、どの画像がユーザにとって必要かなどの判別をする必要があり、その判別基準などについても考察の必要がある。

謝辞

本論文をまとめるにあたり，伊藤哲郎教授には，大分大学在籍時から，多くの御指導，御助言を賜りました．厚く御礼を申し上げます．また，大分大学在籍時および現在九州工業大学においても御指導，御助言を賜りました遠藤勉教授に厚く御礼を申し上げます．

宇津宮孝一教授，川口剛教授，藤田米春教授には，本論文をまとめる上で，多くの御指摘や御助言を賜りました．厚く御礼を申し上げます．

学生時代，日頃から貴重な助言を頂きました大分大学の中島誠助手，賀川経夫助手，伊藤研究室の皆様に深く感謝いたします．

九州工業大学の福本篤史君，林晃司君，森松俊允君には研究を進める上でご協力を頂きました．深く感謝いたします．また，遠藤研究室の皆様にも深く感謝いたします．

両親と妻・直美，そして愛猫・シェルさんにも深く感謝します．

参考文献

- [1] 赤松順子, 高尾宣之, 永井秀利, 中村貞吾, 野村浩郷: 複数製品の紹介記事からの製品情報抽出, 情処学自然言語研報, NL140-9, pp. 61-68, 2000.
- [2] M. Budzikowska et al.: Conversational sales assistant for online shopping, Proc. of HLT2001, 2001.
- [3] J. Chai et al.: A conversational interface for online shopping, Proc. of HLT2001, 2001.
- [4] J. Chai et al.: The role of a natural language conversational interface in online sales: a case study. International Journal of Speech Technology, pp. 285-295, 2001.
- [5] H. H. Chen, S. C. Tsai and J. H. Tsai: Mining tables from large scale HTML texts, Proc. of COLING2000, pp. 166-172, 2000.
- [6] W T. Chuang and J. Yang: Extracting sentence segments for text summarization: A machine learning approach, SIGIR 2000, pp. 152-159, 2000.
- [7] W. Cohen: The Whirl approach to information integration, IEEE Intelligent SYSTEMS, Vol.13, No.5, pp. 20-24, 1998.
- [8] D. Clark: Shopbots become agents for business chance, IEEE Computer, Vol.33, No.2, 2000.
- [9] R. Doorenbos et al.: A scalable comparison-shopping agent for the World-Wide Web, Proc. of the first International Conference on Autonomous Agents, 1997.
- [10] 遠藤勉, 大城英裕, 高岡和宏: 自然言語処理と図形処理を統合化したドリルテキストの意味理解, コンピュータソフトウェア, Vol.8, No.6, pp. 17-27, 1991.

- [11] T. Endo, H. Ohki and K. Takaoka: Cooperative understanding of natural language and picture pattern in drill text, Proc. 1st Conf. of Pacific Association for Computational Linguistics, pp. 131-140, 1993.
- [12] 遠藤勉: 知識ベースの生成を目的としたニュース文の解析, 情報処理学会自然言語処理研究会資料, 58-2,1986.
- [13] T. Endo and T. Kagawa: Cooperative understanding of utterances and gestures in a dialogue-based problem solving system, Computational Intelligence, Vol.15, No.2, pp. 152-169, 1999.
- [14] 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, 情報処理学会論文誌, Vol.38, No.1, pp. 44-54, 1997.
- [15] J. Favela and V.Meza: Image-retrieval agent: integrating image content and text, IEEE Intelligent SYSTEMS, Vol.14, No.5, pp36-39, 1999.
- [16] A. Fukumoto, T. Endo and K. Shimada: Information extraction from specifications on the world wide web, Proc. of PACLING 2001, pp. 109-116, 2001.
- [17] Google: <http://www.google.com/>
- [18] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo: Extracting semistructured information from the Web, Proceedings of the Workshop on Management of Semistructured Data, 1997.
- [19] 林晃司, 嶋田和孝, 遠藤勉: WWW からの性能表抽出のためのキーワード獲得と重み付け, 電子情報通信学会, 思考と言語研究会, 信学技報, 2003.
- [20] R. Houghton: Named Faces: Putting names to faces, IEEE Intelligent SYSTEMS, Vol.14, No.5, pp. 45-50, 1999.
- [21] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong: Medium-independent table detection, Proceedings of Document Recognition and Retrieval VII, pp. 23-28, 2000.
- [22] 飯田仁, 有田英一: 4階層プラン認識モデルを使った対話の理解, 情処学論, Vol.31, No.6, pp. 810-821, 1990.

- [23] 伊藤孝行, 服部宏充, 新谷虎松: エージェント間の協調的入札機構に基づく複数オークション入札支援システム BiddingBot, 人工知能学会論文誌, 人工知能学会, Vol. 17, No. 3, pp. 247-258, 2002.
- [24] 岩根典之, 竹内章, 大槻説乎: 算数の文章題を対象としたネットワーク型知的教育支援環境, 信学論 (D-II), Vol.J80-D-II, No.4, pp. 915-924, 1997.
- [25] 賀川経夫, 高津勝昭, 尾上幸太, 遠藤勉: マルチモーダル対話を用いた問題解決システムにおける音声発話文解析処理, 信学技報, TL98-16, 1998.
- [26] 賀川経夫, 森田光雄, 石橋雄二, 遠藤勉: マルチモーダル対話における発話文生成とジェスチャ生成に関する一考察, 信学技報, PRMU99-156, 1999.
- [27] 金原史和, 佐藤真一, 浜田喬: プリミティブ分解による多様な検索条件を扱うカラー画像検索, 情報処理学会論文誌, Vol. 37, No.11, pp. 1989-2000, 1996.
- [28] 河合敦夫, 塚本雄之, 山本勝紀, 椎野務: 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, pp. 1609-1619, 1998.
- [29] R. I. Kittredge and A. Polguere: The generation of reports from databases, In R. Dale, H. Moisl and H. Somers, eds.): A handbook of natural language processing, pp. 261-304, 2000.
- [30] 熊本忠彦, 伊藤昭, 海老名毅: 支援対話におけるユーザ発話意図の認識 - ユーザ発話文の解析に基づく統計的アプローチ -, 信学論 (D-II), Vol.J77-D-II, No.6, pp. 1114-1123, 1994.
- [31] 串間和彦, 赤間浩樹, 紺谷精一, 木本晴夫, 山室雅司: オブジェクトに基づく拘束画像検索システム: ExSight, 情報処理学会論文誌, Vol.40, No.2, pp. 732-741, 1999.
- [32] 串間和彦, 赤間浩樹, 紺谷精一, 山室雅司: 色や形状等の表層的特徴量にもとづく画像内容検索技術, 情報処理学会論文誌, Vol.40, No.SIG 3(TOD 1), pp. 171-184, 1999.
- [33] 梶井文人, 福本淳一: 製品情報一覧の自動提示のための情報抽出, 言語処理学会第6回年次大会ワークショップ論文集, pp. 56-63, 2000.
- [34] 益岡隆志, 田窪行則: 基礎日本語文法 - 改訂版 -, くろしお出版, 1996.
- [35] 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/index.html>

- [36] 松尾比呂志, 木本晴男: 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, 情報処理学会論文誌, Vol.36, No.8, pp. 1838-1844, 1995.
- [37] 村田剛志, 加藤裕樹, 志村正道, 沼尾正行: 図を用いた算術問題解決システム DIPS, 信学論 (D-II), Vol.J82-D-II, No.1, pp. 75-82, 1999.
- [38] 長尾真: 自然言語処理, 岩波ソフトウェア科学 15, 岩波書店, 1996.
- [39] 那須川哲哉: 頑健な文脈処理のパラダイム - 文脈依存性を考慮した自然言語処理 -, 人工知能学会誌, Vol.11, No.6, pp. 941-949, 1996.
- [40] 日経ベスト PC: 日経 BP 社.
- [41] H.T. Ng, C.Y. Lim, and J.L.T. Koo: Learning to recognize tables in free text, Proceedings of the 37th Annual Meeting of ACL, pp. 443-450, 1999.
- [42] G. Novak Jr. and W. Bulko: Understanding natural language with diagrams, Proc. AAAI80, pp. 465-470, 1988.
- [43] N. Okada and T. Endo: Story generation based on dynamics of the mind, Computational Intelligence, Vol.8, No.1, pp. 123-160, 1992.
- [44] 奥村学, 難波英嗣: テキスト自動要約に関する研究, 自然言語処理, Vol.6, No.6, pp. 1-26, 1999.
- [45] 奥村学, 難波英嗣: テキスト自動要約に関する最近の話題, 自然言語処理, Vol.9, No.4, pp. 97-116, 2002.
- [46] 大平栄二, 阿部正博, 小松昭男, 市川薫: 情報検索における柔軟な対話制御方式, 信学論 (D-II), Vol.J76-D-II, No.3, pp. 586-595, 1993.
- [47] 大津展之, 栗田多喜夫, 関田巖: パターン認識 - 理論と応用 -, 朝倉書店, 1996.
- [48] M. Pazzani and D. Billsus: Learning and revising user profiles: The identification of interesting web sites, Machine Learning 27, pp. 313-331, 1997.
- [49] E. Reiter and R. Dale: Building natural language generation systems, Cambridge University Press, 2000.

- [50] J. J. Rocchio: Relevance feedback in information retrieval, in The SMART Retrieval System: Experiments in Automatic Document Processing, Chapter 14, pp. 313-323, Prentice-Hall, Inc., 1971.
- [51] 桜井成一郎, 志村正道: 算術問題解答システムにおける学習, 人工知能学会誌, Vol.2, No.1, pp. 100-106, 1987.
- [52] しょうがく1年 さんすう塾, 教育書籍, 1989.
- [53] 小1算数5分間トレーニング, 教学研究社, 1990.
- [54] 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, 情報処理学会論文誌, Vol.36, No.10, pp. 2371-2379, 1997.
- [55] 関根聡: テキストからの情報抽出, 情報処理, Vol.40, No.4, pp. 370-373, 情報処理学会, 1999.
- [56] 嶋田和孝, 遠藤勉: 製品性能表からの特徴データの抽出, 情処学自然言語研報, NL133-15, pp. 107-113, 1999.
- [57] 嶋田和孝, 遠藤勉: 特徴化された表データからの要約文生成処理, 信学技報, TL99-29, pp. 25-31, 1999.
- [58] 嶋田和孝, 伊藤哲郎, 遠藤勉: 入力質問文を用いた動的な特徴抽出処理, 信学技報, TL2001-7, pp. 43-50, 2001.
- [59] 嶋田和孝, 伊藤哲郎, 遠藤勉: 表層的な情報を用いた画像の内容特定と分類, 情処学自然言語研報, NL139-8, pp. 55-60, 2000.
- [60] K. Shimada, T. Ito and T. Endo: Classification of images using their neighboring sentences, Proc. of PACLING 2001, pp. 250-256, 2001.
- [61] 嶋田和孝, 伊藤哲郎, 遠藤勉: 表からの特徴抽出処理 - 他種類データへの適用 -, 信学技報, TL2001-33, pp. 27-34, 2001.
- [62] R. Srihari: Computational models for integrating linguistic and visual information: a survey, Artificial Intelligence Review 8, pp. 349-369, 1995.
- [63] R. Srihari: Use of captions and other collateral text in understanding photographs, Artificial Intelligence Review 8, pp. 349-369, 1995.

- [64] J. R. Smith and S. F. Chang: VisualSEEK: A fully automated content-based image query system, Proc. of ACM International Conference on Multimedia, pp. 87-93, 1996.
- [65] M. J. Swain, C. Frankel and V. Athitsos: WebSeer: An image search engine for the World Wide Web, In IEEE Computer Vision and Pattern Recognition Conference, June 1997.
- [66] 高野敦子, 柏岡秀紀, 平井誠, 北橋忠宏: 対話における文脈の定式化と文脈処理の枠組み, 情処学論, Vol.34, No.1, pp. 88-97, 1993.
- [67] 高尾宣之, 永井秀利, 中村貞吾, 野村浩郷: 複数製品の紹介記事からの製品情報抽出 - 製品記述パターンの分析 -, 情処学自然言語研報, NL129-17, pp. 117-124, 1999.
- [68] 田窪行則: 文脈理解 - 文脈のための言語理論, 情報処理, Vol.30, No.10, pp. 1191-1198, 1989.
- [69] 鳥居三朗, 嶋田和孝, 賀川経夫, 遠藤勉: 対話支援型問題解決システムにおける知識獲得方式, 第 49 回電気関係学会九州支部連合大会講演論文集, No.1311, 1996.
- [70] 徳永建伸: 情報検索と言語処理, 言語と計算 5, 東京大学出版, 1999 .
- [71] 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, pp. 77-83, 1996.
- [72] 銀林浩, 相原昭: わかる教え方算数 1 年, 国土社, 1992.
- [73] 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.
- [74] Y. Watanabe, Y. Okada, K. Kaneji and Y. Sakamoto: Retrieving related TV news reports and newspaper articles, IEEE Intelligent SYSTEMS, Vol.14, No.5, pp. 40-44, 1999.
- [75] 渡辺靖彦, 長尾真: 図鑑の解説文から内容抽出を行なうための専門知識の構築, 人工知能学会, Vol.11, No.3, pp. 888-895, 1996.
- [76] 渡辺靖彦, 長尾真: 画像の内容を説明するテキストを利用した画像解析, 人工知能学会誌, Vol.13, No.1, pp. 66-74, 1998.

- [77] M. White et al.: Multidocument summarization via information extraction, Proc. of HLT2001, 2001.
- [78] 山田耕一, 溝口理一郎, 原田直樹: 質問応答システムにおけるユーザ発話モデルと協調的応答の生成, 情処学論, Vol.35, No.11, pp. 2265-2275, 1994.
- [79] 山田剛一, 中川裕志: 人名と顔画像領域との自動対応づけ, 情報処理学会論文誌, Vol.43, No.6, pp. 1890-1898, 2002.
- [80] 山梨正明: 文脈理解への言語学的アプローチ, 人工知能学会誌, Vol.3, No.3, pp. 301-311, 1988.
- [81] 横井俊夫: メディアを手掛かりとした A I 技術・研究の再構築, 人工知能学会誌, Vol.13, No.5, pp. 679-690, 1998.
- [82] M. Yoshida, K. Torisawa and J. Tsujii: Extracting ontologies from World Wide Web via HTML tables, Proc. of PACLING 2001, pp. 332-341, 2001.

索引

- colspan 49
- F 値 44
- rowspan 49
- tf · idf* 87
- インスタンスコマンドフレーム 19
- エントロピー 41
- 応答文の解釈 23
- オンラインショッピング 39
- 格構造 15
- 格要素 15
- 格ラベル 15
- カレントコマンド 21
- 関連項目 56
- 関連属性 57
- 画像の内容推定 82
- キーワード 40, 42, 86
- キーワード候補 40
- 記号メディア 5
- 境界 51
- 行項目 49
- クラスコマンドフレーム 19
- コマンド 10, 18
- コマンドフレーム 18, 21
- 再現率 44
- 最頻度 54
- 修飾要素 15
- 修辞処理 68
- 焦点要素 20
- 情報抽出 38
- 情報統合 3
- スコアリング 54, 57, 58
- 図形構造表現 16
- 性能表抽出 40
- 製品性能表 35
- 製品選択支援システム 36
- 属性間の関連性 57
- 属性の重み 57
- タグの重み 86
- 単位の分類 54, 75
- 知識獲得 32
- 茶釜 40
- 注意状態 20
- 定性的項目 53
- 定量的項目 53
- 適合性判断 59
- 適合性フィードバック 59
- 適合率 44
- データ 49
- 特徴データ 53
- ドリルテキスト 9
- 内容抽出 39
- ノイズワード 42

発話意図	10	要約生成	65
発話文の生成	22	隣接する文章	83
パターンメディア	5	レーダチャート	70
表解析	48	レイアウト解析	85
表構造	49	列項目	49
表構造生成	49		
標準値	53		
表の再構成	69		
表判別のためのキーワード	40		
表領域抽出	43		
— の閾値	43		
フィルタリング	42		
— の閾値	43		
文書の構造	65		
文脈	12		
— 情報	6		
— 処理	1		
— 処理の対象	6		
— の表現	19		
プランニング	21		
平均値	54		
ベイズの定理	42, 88		
法情報	15		
マルチメディア要約処理	92		
マルチモーダル	33		
メディア	5		
— 間の脈絡	6		
— 内の脈絡	6		
メディア間の文脈	82		
問題解決	9		
ユーザの要求	57		
要約	3		