プレゼンテーションスライド情報の構造抽出

羽山　徹彩[†a)]　　難波　英嗣[††]　　國藤　進[†]

Structure Extraction from Presentation Slide Information

Tessai HAYAMA[†a)], Hidetsugu NANBA[††], and Susumu KUNIFUJI[†]

あらまし　近年の電子化プレゼンテーションの普及により，講義や会議などの多くの場面で電子的なプレゼンテーション資料（スライド）が利用され，蓄積されてきた．蓄積されたスライドデータは知識資源として膨大となりつつあるため，その高い利活用性が求められている．スライドデータの利活用性を高めるための効果的な方法の一つとして，レイアウトや視覚的効果など人間の理解を促すための有意な構造情報を利用することが挙げられる．しかしながら，そのような構造情報は，スライドデータの中で明確に定義されていないため，計算機で直接的に扱うことが困難である．そこで，本研究ではスライドに含まれる情報からその構造を抽出する手法を提案する．提案手法は，まずスライド上のオブジェクトを “タイトル”，“図”，“表”，“本文”，“装飾” のいずれかの属性のまとまりに組織化し，それらまとまりをトップダウンに木構造へ組み上げる構造化を行う．評価実験では人手で作成した正解データをもとに，オブジェクトの位置関係に基づいた構造化手法と比較することで，提案手法の有効性を確認した．

キーワード　情報抽出，プレゼンテーションスライド，視覚的レイアウト，Web データ

1. まえがき

近年の電子化プレゼンテーションの普及により，講義や会議などの多くの場面で電子的なプレゼンテーション資料（スライド）が利用されるようになった．利用されたスライドは遠隔講義資料や Web コンテンツとして逐次的に蓄積され，膨大かつ重要な知識資源となりつつある．そのため，スライドに含まれる情報に対して，アクセス性やデータ加工性などの利活用性を高める技術が知識基盤技術として求められている．

スライドに含まれる情報の利活用性を高める有用な方法の一つとして，レイアウトや視覚的情報などの人間の視覚的な理解を促すために情報のまとまりやその関係を表現している有意な構造情報を利用することが挙げられる．しかしながら，これまでのスライドを扱ったシステムのほとんどは，スライドデータを単純なテキストに変換し，キーワードによるアクセス方法をとっており，そのような有意な構造情報を排したデータ管理がなされてきた．このような構造情報を保持したデータ管理ができれば，スライドに含まれる情報をより知的に処理することができるが，構造情報はスライドデータの中で明確に定義されていないため，計算機で直接的に扱うことができない．また，人手により構造情報を付与することは，膨大なコストがかかるため，計算機による自動的な構造情報の抽出が望まれる．

これまで様々なドキュメントを対象とした構造抽出手法が研究されてきた [1], [7], [8]．Rosenfeld ら [6] や Zhai ら [9] は，それぞれ PDF ドキュメントや Web ドキュメントを対象として，機械学習及び木構造テンプレート照合を用いた確率的方法に基づく構造抽出手法を開発してきた．彼らの手法は，構造情報が付与された大量のアノテーション付きデータを必要とし，またその作成された確率モデルが収集データに依存する．そのため，構造パターンが少ないデータを対象に適用することは有効であるが，スライドデータのような多様な構造パターンを含むデータを対象に適用することは難しい．南野ら [5] は，Web ページに含まれる繰返し要素に着目し，Web ページに含まれるテキストの

† 北陸先端科学技術大学院大学知識科学研究科，能美市
Graduate School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1–1, Nomi-shi, 923–1218 Japan
†† 広島市立大学情報科学研究科，広島市
Faculty of Information Sciences, Hiroshima City University, 3–4–1, Ozukahigashi, Asaminami-ku, Hiroshima-shi, 731–3194 Japan
a) E-mail: t-hayama@jaist.ac.jp

構造を抽出する手法を開発してきた．彼らの手法をスライドデータに適用した場合には，HTML タグのような規則性を示す形式的な要素が含まれていないため，そのまま利用することができない．石原ら [3] はスライド音声読み上げシステム構築のために，図に焦点を当てたスライドページ上のオブジェクトの構造抽出手法を開発している．彼らの手法は，オブジェクトの距離関係に基づき，構造情報を抽出している．しかしながら，スライド上のオブジェクトは，自由に作成され，手動で配置されているため，不正確な配置や重なりを避けることができない．そのような場合，オブジェクトの距離関係の利用だけでは，スライドページ全体のオブジェクトの構造情報を適切に抽出することが難しい．以上のように，従来研究ではレイアウトパターン数に限りがあったり，レイアウト内でオブジェクトが正確に配置されていたりするような比較的整ったドキュメント形式をもつデータを対象とし，有効な成果が得られてきたが，それら手法を多彩なレイアウトや不正確なオブジェクト配置を含んだスライドデータに適用することが難しい．

そこで，本研究ではスライドに含まれる情報を対象とした構造抽出手法を開発することを目的とする．本研究で提案する手法は，まずスライドに含まれるオブジェクトを “タイトル”，“本文”，“図”，“表”，“装飾” のいずれかの属性のまとまりに組織化し，それらまとまりをトップダウンに木構造へ組み上げる構造化を行う．

このような構造情報が利用可能になれば，これまでのスライドを利用した様々なアプリケーションの有用性を高めることができる．例えば，スライド音声読み上げシステムではこれまでほとんど利用不可能であった視覚的な構造表現を音声ガイドへ反映させることで，スライド内容をより容易に理解できるような技術が開発可能となる．また，モバイルデバイスなどの小型画面表示領域をもつスライド閲覧システムでは，一度に表示する情報を領域に応じた分割や画面形態に応じたレイアウト割当の技術も開発可能となる．

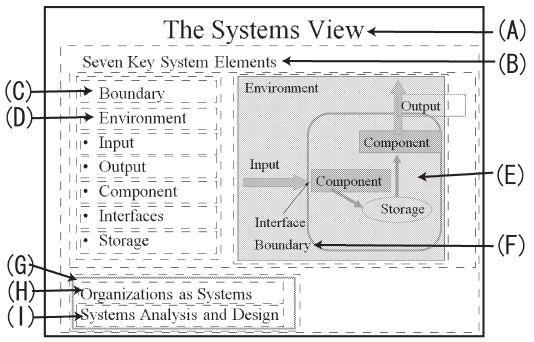
2. スライド情報とその構造

2.1 スライドに含まれる情報

スライドに含まれる情報には，“テキスト”，“写真”，“線” 及び “基本図形” などのプリミティブなオブジェクトから構成されている．本研究におけるスライドに含まれる情報を処理するための前提条件としては，これらプリミティブなオブジェクトのタイプとともに，各オブジェクトのスライド上の縦横位置やフォントサイズの情報が自動的に得られることと，オブジェクトの重なりがあったとしても個々としてオブジェクトを認識できることである．また，インデントや，箇条書き，フォント，表のデータなどの情報は，オブジェクトの位置や情報タイプから判断を行うこととする．このような前提を満たす情報は，Microsoft 社 PowerPoint，Apple 社 Keynote，OpenOffice プロジェクト Impress といった主要なスライド作成ソフトで作成されたスライドファイルにおいて，データとして保存され，XML データとして取り出すことができる．そのため，前提条件となるデータは容易に得ることができる．

このようなプリミティブなオブジェクトは，“タイトル”，“本文”，“図”，“表” 及び “装飾” といったスライド内容を伝える基本表現とするまとまりをなしている．各スライドには，発表の流れに沿ったそのスライドの内容を表現しているタイトルが付与され，そのスライド内容を説明するための項目や補助資料として，本文，図及び表などの基本表現が利用されている．また，それ以外のスライドに含まれているオブジェクトとしては，特定の内容を強調する記号や関係線，あるいは発表日付などのスライド内容と直接関係のない “装飾” 表現がある．このように，スライドに含まれるオブジェクトは，内容に関係する “タイトル”，“本文”，“図”，“表” の 4 種類の属性と，内容に直接関係のない “装飾” 属性のいずれかに分類することができる．

例えば，図 1 の例が示すように，オブジェクト (A)，(C) 及び (F) は，テキストタイプのプリミティブなオ



(点線に囲まれた領域はオブジェクトのまとまりを表現している．)

図 1 スライドに含まれる情報とその構造の例
Fig. 1 An example of slide information and its structure.

ブジェクトであるが，それぞれを “タイトル”，“本文”，あるいは “図” と異なった内容を表現する属性として認識することができる．その際，(F) は (E) やその他のオブジェクトとともに，一つの “図” として内容をもつようなまとまりをなしている．このように，たとえ同じ種類のオブジェクトであっても，異なる属性となったり，単体でなく複数のオブジェクトから組織されたまとまりとなったりすることがある．ここで本論文では，スライド内容を伝える基本表現の性質，及びその基本表現となるオブジェクトのまとまりを，それぞれ機能的属性，及び機能的なまとまりと定義する．

2.2 スライドに含まれる情報の構造

スライドに含まれる情報のもつ構造はスライドの内容を表現するような，オブジェクトの機能的なまとまりを木構造として表現することができる．そのまとまり関係の検出には，スライド上のレイアウトや視覚的効果などに含まれるオブジェクトの位置情報や距離情報を利用することができる．

図 1 の例では，オブジェクト (A) が “タイトル” として機能しており，(A) は木構造の根ノードに相当する．また，周囲のオブジェクトよりも開始位置を下げる字下げは，その前後にあるオブジェクトの階層関係を表現している．その字下げの使用で関係づけられているオブジェクト (B) と (C) は親子ノードとして，また同レベルの箇条書き項目であるオブジェクト (C) と (D) は兄弟ノードとして，それぞれ木構造に割り当てることができる．更に，囲み線 (G) に含まれている複数のオブジェクト (H) と (I) は，(G) が視覚的な閉空間を表現しているため，部分木を構成するとみなすことができる．以上のように，スライド内容を表現する木構造はレイアウトや視覚的効果に含まれる情報を利用することで，主にタイトル属性の機能的なまとまりがその根ノードに割り当て，それと関連する機能的なまとまりをノードとして順次関係づけ，組み上げていくことで構築される．その際，スライド内容に直接関係しない装飾属性のまとまりはその構造に含まれない．

一方，スライド情報の利活用性を高めるための構造情報とは，情報が適切に伝わるようなまとまりとその属性，及びスライドの内容が反映されたそれらまとまりの関係が定義されていることである．その利活用例として，スライド音声読み上げシステムではこれまで各スライド内のオブジェクトの位置順序や作成順序に従って読み上げることを行っていたが，スライド上の内容に関係する情報を適切な分節とそれらの関係を扱えることで，スライドの内容に関する本質的な情報だけを内容に即した順序で読み上げることができる．その結果，ユーザがスライド内容をより正確かつ容易に理解できることが期待される．また，スライド情報検索システムではこれまでスライド上のテキストに対し検索子と一致するスライドの周辺テキストを結果としていたが，情報の属性をもったまとまりとその関係が扱えるようになることで，図/表などの属性を指定したテキスト以外の結果を返す情報検索方法や結果に付随する情報を補助的に提示する情報提示方法への柔軟な拡張が可能となる．

本研究で抽出するスライド情報の構造情報はスライドの内容を伝える基本表現である機能的なまとまりとその属性を特定し，そのまとまりをタイトルをもとにした木構造を組み上げることを抽出することを行うため，スライド情報の利活用性を高めるための構造情報の要件を満たしているといえる．更にスライド情報の利活用性を高めるためには，表理解や図理解，機能的なまとまり関係における修辞構造解析などの意味理解処理を要するが，これら技術を実現するためには大規模な開発が必要となるため，本研究では対象外とする．

3. 提案手法

本研究では，スライドページ上の情報からその構造を抽出する手法を提案する．提案手法は，組織化処理と構造化処理の 2 段階からなる．組織化処理と構造化処理の概要と詳細について，それぞれ **3.1** と **3.2** で述べる．

3.1 組織化処理

図 2 に組織化処理のフローチャートを示す．本手法の組織化では，まず各オブジェクトの属性を “タイトル”，“本文”，“図”，“表” のいずれかに特定し，次に近い距離関係にある同じ属性のオブジェクトをまとめることを行う．オブジェクトの属性特定では，まず各オブジェクトに候補となる属性とその確信度を割り当て，既に属性が確定されたオブジェクトの属性を特定するために影響する他のオブジェクトとの関係を考慮して，より確信度の高いオブジェクトの属性から順次確定していく．ここで本論文では，あるオブジェクトが機能的属性を特定するために影響する他のオブジェクトとの関係を機能的関係と定義する．

属性の種類を確信的に認識できるオブジェクトから優先的に属性特定していくことで，オブジェクトの機能的な属性関係の情報をより正確に扱うことができ，

表 1　属性類ごとの属性らしさを示す得点表
Table 1　Score sheet of attribute based on the likelihood of the attribute.

“タイトル”属性のための評価項目		“本文”属性のための評価項目	
$Ti1$) フォントの大きさ $> Threshold_{(fontsize1)}$	$+1$	$S1$) 箇条書き項目のシンボルがある	$+1$
$Ti2$) トップからの位置 $> Threshold_{(y\_axis\_position)}$	$+1$	$S2$)同じ左位置で同じフォントのテキストタイプのオブジェクトがある	$+1$
$Ti3$) スライド上のオブジェクトの最上位置にある	$+1$	$S3$)左上/右下の位置にテキストタイプのオブジェクトがある	$+1$
$Ti4$) スライドに含まれる中で最大のフォントサイズをもつ	$+1$	$S4$) フォントサイズ $> Threshold_{(fontsize2)}$	$+1$
$Ti5$) 文字数 $> Threshold_{(number\_of\_characters)}$	$+1$	$S5$) 文字数 $> Threshold_{(number\_of\_characters)}$	$+1$
“図”属性のための評価項目		“表”属性のための評価項目	
$F1$) グラフ/画像タイプのオブジェクト	5	$Ta1$) 表に含まれるセルの半数以上にデータが含まれている	5
$F2$)完全にグラフ/画像タイプのオブジェクトと重複している	4	$Ta2$) 表に含まれるセルの半数以下にデータが含まれている	4
$F3$)部分的にグラフ/画像タイプのオブジェクトと重複している	4	$Ta3$)完全に表のセル領域と重複している	4
$F4$)近距離で/間接的にグラフ/画像タイプのオブジェクトと接している	3	$Ta4$)部分的に表のセル領域と重複している	3
$F5$)グラフ/画像の重複したグループの中で最高/最低に位置するテキストタイプのオブジェクト	-1	$Ta5$)近距離で/間接的に表のセル領域と接している	3
$F6$) テキストを含まない基本図形である	4	$Ta6$)表と重複したグループの中で最高/最低に位置するテキストタイプのオブジェクト	-1
$F7$) 文字数 $< Threshold_{(number\_of\_characters)}$	$+1$		

$Threshold_{(fontsize1)}$, $Threshold_{(fontsize2)}$, $Threshold_{(Y_axis\_position)}$ 及び $Threshold_{(number\_of\_characters)}$ は，文字サイズ，文字サイズ，トップからの距離，及び文字数のパラメータを表しており，下線の項目は他のオブジェクトの関係によって評価されることを示している．

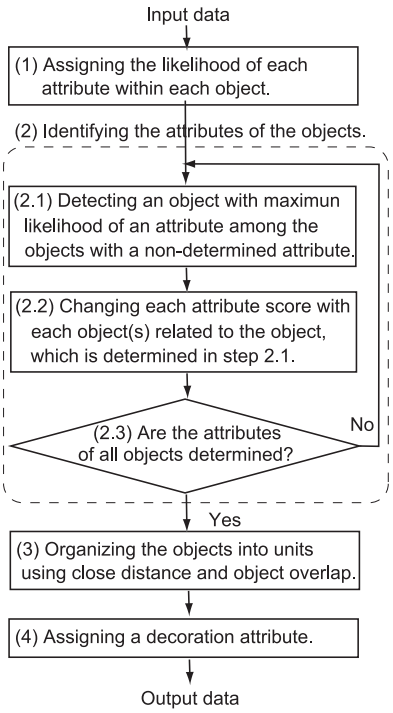


図 2　組織化処理のフローチャート
Fig. 2　Flow chart of organizing processing.

その結果，不確かな属性のオブジェクトに対してもより的確に属性特定することができる．

組織化処理の詳細な手順を以下に示す．

(1)　各オブジェクトの属性類ごとに属性らしさの値を割り当てる

各オブジェクトの候補となる属性とその数値的な確信度を決めるために，各オブジェクトの属性類ごとへ得点付けを行う．オブジェクトの属性類への得点付けには，表 1 の属性類ごとに属性らしさを評価項目とした得点表が利用される．表 1 の各属性類の評価項目の詳細について，以下に示す．

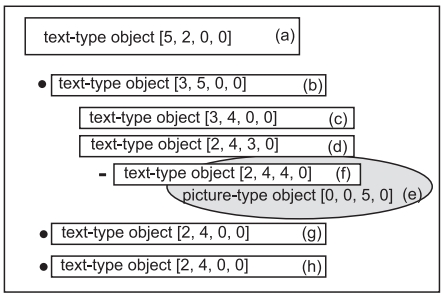
“タイトル” 属性の評価項目：大きなフォントサイズと高い位置にあるオブジェクトに対し，タイトルらしいと考え，高く評価する．規則 $Ti1$ と $Ti2$ ではしきい値より大きいフォントの大きさと高い位置のオブジェクトにそれぞれ加点をし，更に規則 $Ti3$ と $Ti4$ ではスライド内の最高に位置にあるオブジェクトと最大のフォントサイズのオブジェクトにそれぞれ加点をする．また，タイトルはスライドの内容を表現した長さの文字列であると考え，$Ti5$ ではしきい値より長い文字数のオブジェクトに対し，加点をする．

“本文” 属性の評価項目：箇条書き項目の一つ，あるいは周囲の文と開始配置が字下げされたテキストなど，レイアウト構造上で他の本文と関係づけられているオブジェクトに対し，本文らしいと考え，高く評価する．規則 $S1$ と $S2$ では箇条書き項目らしいとして加点をし，規則 $S3$ では周囲の文との間に字下げが適用されているとして加点をする．また，本文は内容があり，見やすい文字列であると考え，規則 $S4$ と $S5$ では文字数の長さ，あるいはしきい値より大きなフォントサイズのオブジェクトに対しそれぞれ加点をする．

“図” 属性の評価項目：グラフや画像のオブジェクトと，それと近距離にあるオブジェクトに対し，図らしいと考え，高く評価する．規則 $F1$ ではグラフと画像のオブジェクトに対し，最大点を付ける．規則 $F2$ と $F3$ ではそれぞれグラフや画像のオブジェクトと重複するオブジェクトに対し得点付けをしており，部分的に重複するよりも完全に重複するオブジェクトに対し高い得点を与える．更に，グラフや画像のオブジェクトとは直接的に重複しないが，近距離に位置してたり，他のオブジェクトを介して間接的に接しているオブジェクトに対して，規則 $F4$ ではそのようなオブジェクトに対し得点を与える．また，以上の重複関係から形成されるグループにおいて最上/最下位置では誤配置されたオブジェクトと重複しやすいと考え，規則 $F5$ ではそのようなオブジェクトの中で図以外の属性となりやすいテキストタイプのオブジェクトに対し，減点をする．規則 $S6$ では図に含まれやすいテキストが単語のような短い文字列であると考え，しきい値よりも文字数の少ないオブジェクトに対し加点をする．

“表” 属性の評価項目：格子状の囲み線とその囲みに位置するオブジェクトに対し，表らしいと考え，高く評価する．規則 $Ta1$ と $Ta2$ では格子状の囲み線において表のセルデータが占められている方が表らしいと考え，表の格子中にデータが多く満たされている格子状の囲み線となるオブジェクトに対し高い得点を与える．規則 $Ta3$ では表に含まれるデータとして，格子状の囲みと重複しているオブジェクトに対し，得点を与える．更に，表データへの注釈も表の一部であるとみなし，規則 $Ta4$ と $Ta5$ では格子状の囲み線の領域と部分的に重複，あるいは近距離に位置したり，他のオブジェクトを介して間接的に接しているオブジェクトに対し，それぞれ得点を与える．また，誤配置されたオブジェクトが表とその重複するオブジェクトのグループにおいて最上/最下位置で重複しやすいと考え，規則 $T6$ ではそのようなオブジェクトの中で表以外の属性となりやすいテキストタイプのオブジェクトに対し，減点をする．

オブジェクトの属性類ごとの得点付けでは，適合する評価項目の総得点が割り当てられる．その際，他のオブジェクトと関係づけることで属性らしさを評価する項目（表 1 の下線項目）が適用された場合には，その関係したオブジェクトを属性類ごとにリスト化する．本論文では，そのリストを機能的関係リストと呼ぶこととする．



The numbers within square brackets indicate the attribute scores of title, body text, figure and table.

図 3　属性得点が含まれるスライドの例
Fig. 3　An example of a slide including attributes scores.

オブジェクトの属性類ごとの得点付けの例を図 3 に示す．Object(b) の属性類 [“タイトル”，“本文”，“図”，“表”] には，[3, 5, 0, 0] の得点が付けられる．その際，Object(b) の “本文” 属性の機能的関係リストには，Object(c)，(g) 及び (h) が含められる．

(2)　オブジェクトの属性を決定する

(1) で設定された属性類ごとの属性らしさの値を利用することで，各オブジェクトの候補となる属性とその確信度を算出し，その確信度が高いオブジェクトから順に他のオブジェクトとの機能的関係を考慮しながら属性を確定していくことで，すべてのオブジェクトの属性を決定する．

その手順の詳細を，(2.1) から (2.3) に示す．

(2.1)　属性が未確定のオブジェクトの中から，その候補となる属性の確信度が最も高いオブジェクトを選出し，その属性を確定する．

はじめに，まだ属性が確定されていない各オブジェクトに対し，“タイトル”，“本文”，“図”，“表” の四つの属性類の中で得点が最も高い属性類の一つを候補となる属性とする．次に，それらオブジェクトの候補となる属性の確信度を算出する．属性の確信度は，その属性らしさが高いだけでなく，その他の属性類の項目において属性らしくなさも考慮する必要がある．そこで，候補となる属性の確信度（Li_Attri）はその両方の性質を考慮した式 (1) と (2) によって算出される．

$$Ev_{(attri)} = \begin{cases} Attri\_Val_{(attri)} \\ (if\ attri\_cand == attri) \\ MaxScore_{(attri)} \\ \qquad - Attri\_Val_{(attri)} \\ (otherwise) \end{cases} \quad (1)$$

$$Li\_Attri = Ev_{(`title')} * Ev_{(`body-text')} * Ev_{(`figure')} * Ev_{(`table')}. \quad (2)$$

ここで，$attri$，$Attri\_Val_{(attri)}$，$attri\_cand$ 及び $MaxScore_{(attri)}$ は，ある属性とそれに付けられた得点，候補となる属性及び属性類ごと最大得点(注1)を示している．式 (1) の $Ev_{(attri)}$ は，$attri$ が候補となる属性である場合にその属性に付けられた値をとり，$attri$ がそれ以外の属性類である場合にその属性の最大得点からその属性に付けられた得点を引いた値，つまり属性らしくなさの値をとる．次に，式 (2) の Li_Attri は各オブジェクトにおいて式 (1) で得られたすべての属性の値を積算した結果となる．その結果では，候補となる属性の得点が高く，それ以外の属性類の得点が低い場合に，確信度が高い値となる．一方，候補となる属性の得点とそれ以外の属性類の得点が拮抗していた場合には，確信度が低い値となる．これらの式を用いて，属性が未確定なオブジェクトの中で確信度が最大のオブジェクトに対し，その候補となる属性を属性として確定する．また，その確信度が最高のオブジェクトが複数ある場合には，その中でスライド上の上位置にあるオブジェクトに対し，属性を確定する．

図 2 の例において，Object(b) と (g) の候補となる属性はともに “本文” 属性となり，その確信度にはそれぞれ 375 及び 300 が算出される．その結果，object(b) は object(g) よりも候補となる属性の確信度が高いため，優先的に属性が確定される．

(2.2)　新たなオブジェクトの属性確定に伴い，その機能的関係の影響を他のオブジェクトへ与える．

(2.1) で新たに確定されたオブジェクトに対して，その属性以外の属性の得点付けで機能的関係にあると判断されたオブジェクトとの関係は不適切である．そのため，そのオブジェクトの属性以外の属性の機能的関係リストに含まれているオブジェクトに対し，その属性らしさの値を再計算するとともに，確定されたオブジェクトを機能的関係リストから取り除くことを行う．また，各ページのタイトルを唯一とするために，そのオブジェクトが “タイトル” 属性と確定されたなら，その他のオブジェクトの “タイトル” 属性らしさの値を 0 に設定する．

図 3 の例では，object(d) が “本文” として属性確定されたなら，object(d) の “図” 属性の機能的関係リストに含まれている object(f) の “図” 属性らしさの値は 3 に再設定される．また，object(a) が “タイトル” として属性確定されたなら，その他のオブジェクトの “タイトル” 属性らしさの得点は 0 に再設定される．

(2.3)　スライドページ上のすべてのオブジェクトの属性が特定されるまで，(2.1) と (2.2) の手順を繰り返す．

(3)　距離関係に基づきオブジェクトを組織化する

すべてのオブジェクトの属性が特定された後，“図”/“表” 属性のオブジェクトに対し，“図”/“表” 属性の機能的関係リストに含まれるオブジェクトを一つにまとめる．その際，(2.2) で属性確定されたオブジェクトに関連する他のオブジェクトの機能的関係リストも更新されているため，同じオブジェクトが異なる “図”/“表” 属性の機能的なまとまりに含まれることがなく組織化される．

(4)　装飾属性を割り当てる

本文を内包する基本図形や図に含まれない矢印図形は，オブジェクトを明示的に関係づける表現として使用されるため，内容と直接関係のない装飾とみなすことができる．そこで，“本文” 属性のまとまりを内包している基本図形のオブジェクトといずれのまとまりにも組織化されていない矢印図形のオブジェクトに対し，“装飾” 属性を割り当てる．

3.2　構造化処理

本手法の構造化では，トップダウンによる領域分割に基づいた方法で行う．つまり，オブジェクトの機能的なまとまりを含む領域を段階ごとに分割していき，各分割段階を親子ノードとして関係づけていくことで階層構造を得ることができる．この領域分割では，スライドページに含まれる視覚的なレイアウト構造の規則性を検出し，利用する．またレイアウト構造の規則性の検出が難しい場合には，各領域に含まれるまとまりの属性の並びによって，領域分割の位置を判断する．視覚的な位置だけでなく，異なる属性の並びの規則性も利用することで，位置関係だけに依存しない領域分割が可能となり，不規則なレイアウト構造に対しても柔軟に対応することができる．

(注1)：表 1 の得点表では，すべての属性類の最大得点が 5 である．

今回の構造化処理では，対象データが Web から収集された情報科学技術分野の発表資料を多く含んでいたため，横書きを基本とした方法となっている．そのため，横書きを基本としたスライドのレイアウト構造はページをブロック単位に分ける段組みが縦方向の分割点をもつため，本構造化処理の手順では，まず縦方向への領域分割を試みてから，横方向への領域分割を行う．

構造化処理の詳細な手順を以下に示す．

(1)　初期設定

領域分割を行うための初期領域と木構造の根ノードを設定する．スライドページに “タイトル” 属性のまとまりを含んでいるなら，根ノードと初期領域にはそれぞれそのまとまりとそのまとまり以外のページ領域が割り当てる．一方，“タイトル” 属性のまとまりが含まれていないならば，根ノードは空ノードとし，初期領域にはページ全体が割り当てる．

(2)　縦方向への領域分割

領域内に縦断する空領域が含まれているなら，その領域は空領域によって分割される．

(3)　横方向への領域分割

領域処理の操作では領域内の左上に位置するまとまりを基準として，レイアウトの規則性や属性の並びを調べることで，異なる条件によって分割を行う．その分割条件として以下の三つのうち，いずれか一つの条件が適用される．その際，領域全体を占めている “装飾” 属性の囲み記号が複数の機能的なまとまりを囲んでいる場合にはこれ以上の領域分割処理を進めることができないため，その囲み記号をまず除外してから，分割条件の適用を行う．それによって，“装飾” 属性の囲み記号に内包されている複数の機能的なまとまりに対し，部分木となるように構造化することができる．

三つの分割条件の詳細を以下に述べる．

分割条件 1：領域内に横断する空領域を検出する．もし，その空領域が指定したしきい値以上の分割幅であるならば，領域はその空領域によって分割される．

分割条件 2：領域内のまとまりの属性を調べる．もしその属性の並びが “本文” 属性と “図”/“表” 属性との関係からなる以下の規則に適合するなら，その領域は各規則に従って分割される．各規則について，領域に含まれるまとまりの属性の並びとその分割位置を示した図 4 をもとに説明する．block(a) 及び (b) は，領域内の最上位置に “本文” 属性のまとまりがあり，更に (a) には最上位置にあるまとまりよりも左に位置する “図”/“表” 属性のまとまりがあるが，(b) には最上位置にあるまとまりよりも左位置に他のまとまりがない例である．block(d) は，領域内の最上位置にあるまとまりが箇条書き項目とする “本文” 属性である例である．また，block(c) は，領域内の最上位置にあるまとまりが “図”/“表” 属性の例である．

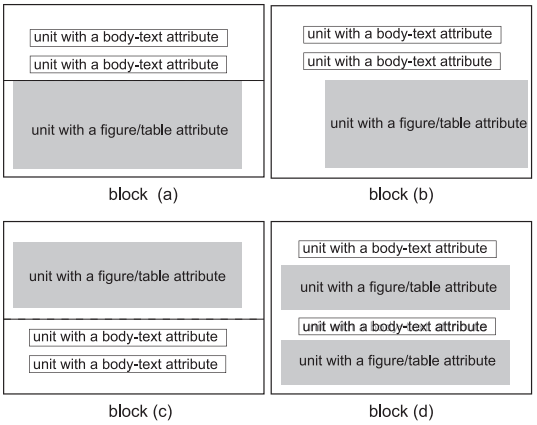


図 4　領域に含まれるまとまりの属性の並びとその分割位置

Fig. 4　Units’ attribute sequence in a block and it’s dividing point.

(i)　領域内の最上位置にあるまとまりが “本文” 属性であり，そのまとまりよりも左位置にある “図”/“表” 属性のまとまりが含まれているなら，その領域はその “図”/“表” 属性のまとまりの上位置で分割される．ただし，その “図”/“表” 属性が箇条書き項目の間に位置する場合は除く．この規則によって，block(a) には適用され，破線位置で分割されるが，block(b) と (d) には適用されない．

(ii)　領域内の最上位置にあるまとまりの属性が “図”/“表” 属性であるなら，その領域はその “図”/“表” 属性のまとまりの下位置で分割される．この規則によって，block(c) には適用され，破線位置で分割される．

分割条件 3：領域内の左上位置の機能的なまとまりを調べる．もし，そのまとまりが箇条書き項目に含まれている “本文” 属性であるなら，その領域はその箇条書きの各項目の上位置で分割される．もし，そのまとまりが箇条書き項目に含まれない “本文” 属性であるなら，その領域はそのまとまりとそれ以外に分割される．

(4)　すべての領域に対して，まとまりがたかだか一つ含まれるまで，(2) と (3) の分割処理を繰り返す

4. 評価実験

4.1 概　　要

我々は提案手法の有効性を明らかにするために，以下の点に焦点を当てて，評価実験を実施した．

- 組織化において，オブジェクトの距離関係とともに機能的関係の情報を用いることの有効性
- 構造化において，視覚的な手掛りの規則性とともに属性関係の規則性を用いることの有効性

これまでスライド上の情報を対象とした構造抽出手法やそのための評価データは存在しないため，我々は比較手法とその評価データを作成した．まず組織化の比較では，距離関係の情報だけを利用した方法を用いた．その具体的な処理としては，“図”/“表” タイプのオブジェクトと重複や近距離に位置するオブジェクトに対し，“図”/“表” 属性のまとまりとして組織化することを行った．次に構造化の比較では，視覚的な手掛りの規則性だけを利用した方法を用いた．その具体的な処理としては，レイアウトや視覚的効果に含まれるまとまりを以下の関係づけ表現に基づいた規則によって，トップダウンに領域分割を行った．

・“タイトル” 属性のオブジェクトを根ノードに割当
・字下げされたまとまりとその直前のまとまりを親子関係のノードとする
・同じレベルの箇条書き項目や左位置がそろっているまとまりを兄弟関係のノードとする
・“装飾” 属性の囲み記号で内包されたまとまりを部分木として扱う

組織化の評価方法として，$Precision$，$Recall$，及び $F-measure$ の指標が利用された．その値は以下の (3)〜(5) の式で算出される．

$$Recall = \frac{Matched\_CorrectData}{Total\_CorrectData} \tag{3}$$

$$Precision = \frac{Matched\_CorrectData}{Total\_DetectedData} \tag{4}$$

$$F-measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5}$$

ここで $Matched\_CorrectData$,$Total\_CorrectData$, 及び $Total\_DetectedData$ は，正解データとの適合数，正解データの総数，及び検出データの総数を示す．また，構造化の評価では，各スライドページ内でのまとまりの関係づけの正確さによって比較した．

評価データとその正解データには，Web からの自動収集データを含むデータベース [4] から 98 組の日本語スライドデータをランダムに選択し，利用した．そのデータの平均ページ数は 24.14 ページであり，総ページ数は 2366 ページとなる．正解データの作成には属人性の影響を配慮し，スライドの閲覧することに慣れた作成者の選定と，項目説明と手順のマニュアル化を行った．正解データの作成者は，7 回以上の学会発表経験をもつ博士課程の大学院生 2 人が選ばれ，独自に開発した編集ツールを使用し，オブジェクトの機能的なまとまりとその属性，及び構造関係の定義付けを行った．その際，構造の識別が難しい場合には，無理な定義付けを与えないようにした．作成者への事前指導ではサンプルとして 5 種類の正解データ（平均 18 枚のスライド）を与え，まず属性類の意味の説明を行い，次に作成手順として，1）属性類を指定したオブジェクトのまとまりを作成，2）タイトルをもとにしたまとまり同士の関係付け，3）関係づけられないまとまりを “装飾” 属性と同定，に従って実施するように説明が与えられた．

本実験では，提案手法と比較手法を実装した実験システムが用いられた．実験システムは，スライドファイルから自動的に各ページに含まれているオブジェクトを抽出し，構造抽出処理が実行され，その結果としてオブジェクトのまとまりやその属性，及び構造に関する情報をメタデータとした XML 形式のファイルが出力される．現在のシステムは，Microsoft Visual Studio C#によって実装され，Microsoft PowerPoint（PPT）ファイルを入力としている．我々は PPT ファイルのオブジェクト抽出において，オブジェクトとその情報タイプと位置，フォントサイズの情報だけを使用し，PPT ファイルデータに含まれるレイアウトテンプレートの論理構造の情報に対しては，必ずしもレイアウトの規定に沿ったデータ入力となっていないため，利用しなかった．

システムの出力例として図 1 のスライドを入力した結果を図 5 に示す．図 5 では，まとまりを示すタグ（“Unit”）にオブジェクトを示すタグ（“Object”）が内包されることで機能的なまとまりを表現しており，また各まとまりの属性は属性を示すタグ（“attribute”）に含まれている．各スライドページに含まれる情報の構造は，それらまとまりの関係を示すタグ（“NodeList”）に，まとまり番号（“Unit ID” タグに含まれる数）によって関係づけられている．

本実験で用いた提案手法のパラメータ設定値は表 2 のとおりである．

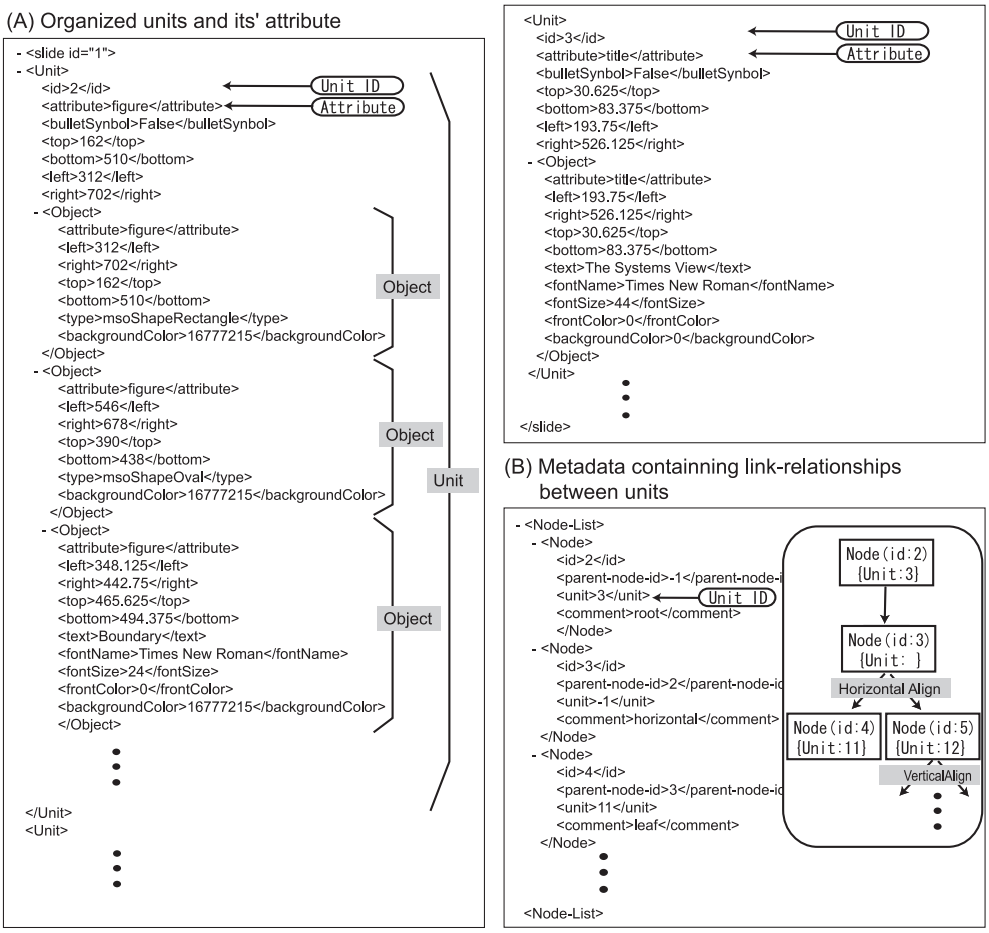


図 5　提案手法に基づいて構築された実験システムによる XML データの出力例
Fig. 5　An example of XML data outputted by an experimental system developted based on proposal method.

表 2　本実験で使用した提案手法のパラメータ
Table 2　Parameters of proposal method used in this experiment.

パラメータ	値
$Threshold_{(fontsize1)}$	24 pt
$Threshold_{(fontsize2)}$	32 pt
$Threshold_{(Yaxis\_position)}$	スライドの縦 1/4 のサイズ
$Threshold_{(number\_of\_charactors)}$	8 文字
構造化処理の分割条件 1 の幅	24 pt

4.2　結果と考察

組織化と構造化を行った実験結果を，それぞれ表 3 と表 4 に示す．

表 3 は，オブジェクトのまとまりとその属性の正確さを属性ごとに分類した評価結果である．表 3 が示すように，提案する組織化手法は比較手法よりも，すべての属性において精度が高かった．特に，“図” 属性のオブジェクトのまとまり検出では，$F-measure$ が提案手法 0.89 に対し比較手法 0.69 と，顕著に効果的であることが確認された．“図” 属性のオブジェクトのまとまりは，重なりや近さの距離情報によってまとまりを構成されることが多いため，不適切なオブジェクトの配置に影響を受けやすい．そのため，提案手法で用いているオブジェクトの機能的関係の情報を利用することが，オブジェクトの不適切な配置を検出し，適切な属性へ割り当てることに有効であったといえる．

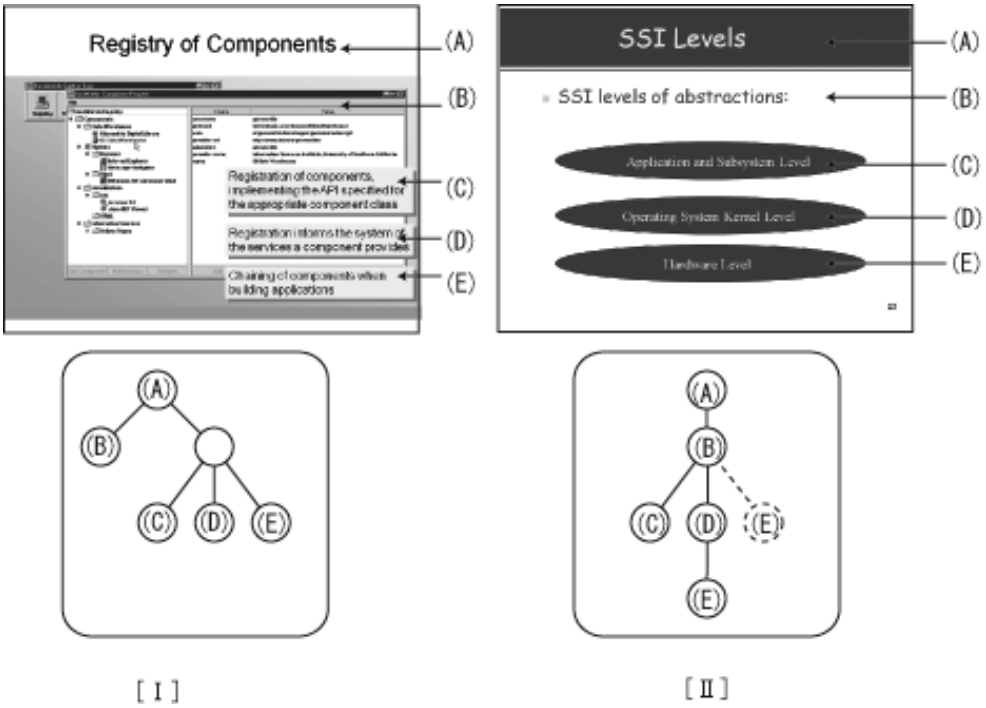
表 4 は，各ページにおいてまとまりを関係づけた精度ごとの割合を表した結果である．表 4 が示すように，提案する構造化手法は比較手法よりも，それらまとまりを完全に関連づけられている割合が 0.95 に対し 0.90 と高かった．そのため，構造化手法では不規則

表 3　組織化処理の属性ごとの精度
Table 3　Accuracy for each attribute results in the organizing process.

属性類とそれらまとまりの正解データ数		タイトル (2333)	本文 (9285)	図 (1905)	表 (46)	装飾 (2201)
提案手法	Recall	0.97	0.89	0.93	0.96	0.96
	Precision	0.99	0.85	0.85	0.98	0.81
	F-measure	0.98	0.85	0.89	0.97	0.87
比較手法	Recall	0.87	0.69	0.64	0.93	0.91
	Precision	0.96	0.88	0.63	0.93	0.63
	F-measure	0.92	0.77	0.64	0.93	0.74

表 4　構造化処理におけるページ内のまとまりの関連付け精度の割合
Table 4　Ratio in pages for each correct ratio of results in the structuring process.

ページ内のまとまりの関連付け精度の範囲	1.00	0.99〜0.80	0.79〜0.60	0.59〜0.00	N/A
提案手法（組織化：提案手法）	0.95	0.03	0.04	0.05	0.12
比較手法（組織化：提案手法）	0.90	0.05	0.06	0.07	0.12
比較手法（組織化：比較手法）	0.76	0.07	0.08	0.15	0.12



(点線で囲まれたノードは正解データにおける位置を表している．)

図 6　本構造化手法の抽出結果が正解データと一致した例 [I] と一致しなかった例 [II]
Fig. 6　Slide samples matching/mis-matching structure data extracted by the proposal method to its' correct data.

なレイアウトを補うために，属性関係の規則性を利用することが有効であるといえる．更に，提案する一連の構造抽出手法の特徴としては，属性を特定し，その情報を利用することが挙げられる．本実験結果において，一連の提案手法の適用によって完全に構造抽出できる割合は 0.95 であり，属性情報を用いない比較手法の 0.76 に比べ，大幅な向上が見られた．そのため，スライドに含まれる情報の構造抽出には，属性情報を利用することが有効であるといえる．

我々は実験結果より，提案手法が引き起こした主なエラーの原因を確かめた．その原因の一つは，オブジェクト間の関係を視覚的な構造で表現するのでなく，テキストの記述内容で定義されている場合ある．例えば，図のオブジェクトとその説明テキストが切り離さ

れた位置にあり，記号などで対応付けされている場合がある．そのような原因に対し，オブジェクトの構造関係を適切に検出するためには，簡単なテキスト分析を行う必要がある．

また提案手法は横書きを基本とした研究発表スライドをもとにして本手法のルール群が作成されているおり，本実験結果から本手法の様々な適用制限が明らかとなった．まず，中ぞろえ，あるいは右ぞろえの箇条書き項目が含まれていた場合には，各項目の左開始位置が異なるため字下げの使用と判断されることもあり，それが不適切な構造化へ導くこととなる．また，縦書きと横書きが混在している場合には，それらを正確に構造化することができない．このような場合の対処方法としては前処理として，箇条書き項目のそろえ位置や横/縦書きの判断を行うことで，そのためのルールを適用する必要がある．以上のようなエラー原因は横書きを基本とするルールの適用によるものであるが，本実験において 95%の精度で構造抽出が可能であるため，まれな場合であるといえる．

最後に，本構造化手法の抽出結果が正解データと一致した例 [I] と一致しなかった例 [II] を図 6 に示す．

本構造化手法の抽出結果が正解データと一致した例では，“本文” 属性のオブジェクト (C)，(D)，(E) が図 (B) と重なっていたにもかかわらず適切に構造情報を抽出することができていた．また，本構造化手法の抽出結果が正解データと一致しなかった例では，正解データにおいて “本文” 属性のオブジェクト (C)，(D)，(E) が兄弟関係ノートとして構造化されていたにもかかわらず，中ぞろえとなっていたため，(E) が (D) に対して字下げされていると判断され，親子関係ノードとして構造化されていた．

5. むすび

本論文では，膨大かつ重要な知識資源となりつつあるスライドデータの利活用性を高めるための基礎技術として，スライドページに含まれる情報の構造抽出手法を提案した．提案手法では，まずスライドに含まれるプリミティブなオブジェクトを機能的なまとまりへ組織化を行い，それらまとまりをトップダウンに木構造へ組み上げる構造化を行う．その際，組織化ではスライド上のオブジェクトの不正確な配置や重なりに対処するために，距離関係だけでなく機能的な関係に関する情報を利用した．また，構造化ではレイアウトの規則性が損なわれる問題に対し，例外的な対応に属性関係の規則性を利用した．評価では人が作成した正解データをもとにした比較実験により，提案手法の有効性が確認された．

現在のシステムはまだ改善が必要であるが，本実験結果からスライドに含まれる情報の構造抽出を 95%の精度の正確さで可能であることが分かった．今後は，提案手法で抽出された構造情報をメタデータとして利用することで，スライドデータを扱った様々な技術を開発していきたい．その一例として，構造情報を利用したスライド中の情報の検索システムや，レイアウト構造を変換することで検索結果の複数スライドページを分かりやすい形で提示する閲覧インタフェースの開発が挙げられる．また，スライド上の言語表現分析技術 [2] の開発も行っていきたい．

謝辞　本研究成果の一部は，財団法人電気通信普及財団 平成 21 年度研究調査助成金，及び科研費（基盤研究 B，20300046）の助成により実施されたものである．

文　献

[1] A. Anjewierden, “AIDAS: Incremental logical structure discovery in PDF documents,” Proc. 6th International Conference on Document Analysis and Recognition, pp.374–378, 2001.

[2] T. Hayama, H. Nanba, and S. Kunifuji, “Alignment between a technical paper and presentation sheets using a hidden Markov model,” Proc. Active Media Technology 2005, pp.102–106, 2005.

[3] T. Ishihara, H. Takagi, T. Itoh, and C. Asakawa, “Analyzing visual layout for a non-visual presentation-document interface,” Proc. 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp.165–172, 2006.

[4] H. Nanba, T. Abekawa, M. Okumura, and S. Saito, “Bilingual presri: Integration of multiple research paper databases,” Proc. 7th RIAO Conference: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp.195–211, 2004.

[5] 南野朋之，斎藤　豪，奥村　学，“繰返し構造に基づいた Web ページの構造化，” 情処学論，vol.45, no.9, pp.2157–2167, 2004.

[6] B. Rosenfeld, R. Feldman, and Y. Aumann, “Structural extraction from visual layout of documents,” Proc. 11th International Conference on Information and Knowledge Management, pp.203–210, 2002.

[7] T. Watanabe, Q. Luo, and N. Sugie, “Layout recognition of multi-kinds of table-form documents,” IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.4, pp.432–445, 1995.

[8] Y. Yang and H. Zhang, “HTML page analysis based

on visual cues," Proc. 6th International Conference on Document Analysis and Recognition, pp.859–864, 2001.
[9] Y. Zhai and B. Liu, "Structured data extraction from the Web based on partial tree alignment," IEEE Trans. Knowl. Data Eng., vol.18, no.12, pp.1614–1628, 2006.
（平成 20 年 12 月 15 日受付，21 年 4 月 13 日再受付）

羽山　徹彩　（正員）

2001 同志社大・工・知識工学卒．2003 北陸先端科学技術大学院大学知識科学研究科博士前期課程了．2006 同大学院大学知識科学研究科博士後期課程了．同年北陸先端科学技術大学院大学知識科学研究科助手．2007 助教．博士（知識科学）．現在は主として，知識システム，創造性支援システム，ヒューマンインタフェースの研究に従事．人工知能学会，情報処理学会，日本創造学会各会員．

難波　英嗣

1996 東京理科大・理工・電気卒．1998 北陸先端科学技術大学院大学情報科学研究科博士前期課程了．2001 同大学院大学情報科学研究科博士後期課程了．同年日本学術振興会特別研究員．2002 東京工業大学精密工学研究所助手．同年広島市立大学情報科学部講師．2007 広島市立大学大学院情報科学研究科講師，現在に至る．博士（情報科学）．テキストマイニング，情報検索，自動要約，特許情報処理に関する研究に従事．言語処理学会，人工知能学会，ACL，ACM 各会員．

國藤　　進　（正員）

1974 東京工業大学理工学研究科修士課程了．同年（株）富士通国際情報社会科学研究所入所．1982〜1986 ICOT 出向．1992 北陸先端科学技術大学院大学情報科学研究科教授．1998 知識科学研究科教授．現在は主として発想支援システム，グループウェア，知識システムの研究に従事，情報処理学会創立 25 周年記念論文賞．人工知能学会 1996 年度研究奨励賞，日本創造学会 2004 年論文賞などを受賞．博士（工学）．情報処理学会，計測自動制御学会，日本創造学会等各会員．