Vol.2010-DD-74 No.6
2010/1/29

セグメントオーバーレイ方式による プレゼンテーション資料の構造化と 社内文書検索への応用

山本康高† 松田勝志††

プレゼンテーション用の資料は、文書管理に役立つ節の階層性が把握しづらい。従来技術では、目次がある、各スライドのタイトルに番号があるなど、目次に相当する情報が明記されているプレゼンテーション資料しか節の階層性を推定できなかった。本稿では、トピックとなるスライドの集合を再帰的に抽出する方式を提案する。実験により、提案方式が、目次に相当する情報がないプレゼンテーション資料からでも、各スライドの節の階層性を約 60%の正答率で推定できること、推定した構造が社内情報検索の効率化に寄与することを確認した。

Development of Hierarchical Document Structure Inference for Presentation Materials and Application to Enterprise Search

Kosuke Yamamoto† and Katsushi Matsuda††

This paper presents segment-overlay method which can infer a hierarchical document structure of presentation materials. The segment-overlay recursively subdivides a document into segments by detecting characteristic slides from the document, overlays partial hierarchical structures among the segments and generates the document structure. This paper shows that segment-overlay can infer the document structure with about 60% accuracy, and also reports that inferred document structure makes users easier to search presentation materials.

1. はじめに

複数のトピックからなるオフィス文書をトピック単位で管理・検索できれば，文書の再利用を促進でき，多くの企業における文書の生産性の向上を支援できる．特に，本稿では，表やグラフなど作成に手間がかかるコンテンツを多く含むプレゼンテーション資料（プレゼン資料）に着目する．

文書は，一般に節によってトピックがまとめられるため，節の構造により文書内のトピックを大まかに分類できる．

そこで，我々はプレゼン資料のトピックの分類を節の構造（目次構造）を推定する問題と捉える．目次構造の推定とは，プレゼン資料の各スライドが属する節（1 節や 1.1 節など）を決めることである．これにより，プレゼン資料をトピック単位で再利用できるだけでなく，一つのトピックにかけるスライドの枚数などで，作成者が注力したいトピックも分析できるようになる．

プレゼン資料の目次構造を推定する従来技術もあるが，それらは下記に示す前提を設けているため，目次構造を推定できるプレゼン資料が限られている．

- プレゼン資料内に目次が記載されたスライドがある，もしくは各スライドのタイトルに「1. Introduction」など節の番号が記録されているなど，明記された目次に相当する情報（以降，目次相当情報）がある．
- ある特定の文字列を含むスライドを第 1 節とみなすなど，スライドの特徴と節とが 1 対 1 に対応する．

そこで，任意のプレゼン資料から目次構造を推定することを本研究の課題とする．

本稿で提案するセグメントオーバーレイは，一つのトピックとなるスライドの集合（以降, セグメント）の抽出を再帰的に行うことで目次構造を推定する. 提案手法は，セグメントを複数の特徴的な書式のスライドを検出し決めるため，目次相当情報を必須としない．また，セグメントの抽出を再帰的に行い節の階層性を特定するため，事前にスライドの特徴と節とを 1 対 1 に対応させる必要がない．

実験では，目次相当情報がないプレゼン資料から約 60%の正答率で目次構造が推定できることを示す．また，提案手法で推定した目次構造が，プレゼン資料の検索の効率化に寄与することを示す．

2. 目次構造

本稿では，目次構造を，各スライドが属する節を推定して得られるプレゼン資料全

† 日本電気株式会社 共通基盤ソフトウェア研究所
Common Platform Software Research Laboratories, NEC Corporation
†† NEC ビッグローブ
NEC Biglobe

Vol.2010-DD-74 No.6
2010/1/29

体の木構造と捉える．プレゼン資料の全体を代表するスライドを表紙として，目次構造を模式的に表した例を図 1 に示す．図 1 の「#数字」はスライドのページ数，「§数字」は節の番号を表す．図 1 の目次構造は，大きく分けて#2～#4 と#5～#7 の 2 つの節から構成され，各節はさらに 2 つの節を有する．このとき，#2，#5 でトピックが切り替わること，ならびに，各スライドの階層の深さを求めることが目次構造推定となる．図 1 の#2，#3 が§1 や§1.1 であると決定できる情報がプレゼン資料中に明記されているとは限らない．

しかしながら，論理的かつ分かりやすい資料の作り方には多くの作成者に共通するノウハウがある．逆にそれらは目次構造を推定するための情報源になり得る．ノウハウの例を以下に示す．これらのノウハウは，我々が 2 万人規模の営業用支援ツールに蓄積されているプレゼン資料を見，多くの人に共通して見られる汎用的なものである．

- タイトルのみを記載したスライドを途中に挿入しスライド間の区切れ目を明確にし，節の区切れ目を分かりやすくする．
- 説明する内容を予めリストで列挙し，列挙された内容を後のスライドで詳述することで，スライド間に意味的な階層性を形成する．

目次相当情報に加え，これらのノウハウを目次構造の推定に利用する．

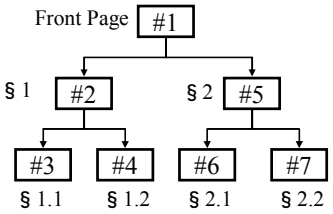

図 1　プレゼンテーション資料の目次構造

1 つの文書を複数のトピックに分類する技術分野は Text/Topic segmentation などと呼ばれる．その多くは，単語の共起性などに基づいて節を切り出す [1)2)]．また，統計的解析に基づく手法 [3)]，HMM により単語の推移を考慮する手法 [4)]，教師データを必要としない構造化手法 [5)]なども提案されている．しかしながら，これらはテキストの分割が目的であり，その階層性までは考慮しない．三木 [6)]は，プレゼン資料を対象とし，キーワードの類似性によりスライドをまとめることで節を特定しているが，階層性の推定は行えない．これらの手法を再帰的に適用し，節を段階的に分割／統合し目次構造を推定することも考えられるが，テキスト情報の少ないプレゼン資料に対して，テキストに基づく手法の適用は不向きである．その他，Klink[7)]らのようにレイアウトに基づき，文書を構造化する手法もあるが，プレゼン資料は同一のレイアウトのスライドも多く，レイアウトの情報だけでは目次構造の推定は困難である．

3. セグメントオーバーレイ

本稿では，目次構造の推定を，各スライドの節の階層の深さと，トピックの起点となるスライドを推定する問題と捉える．セグメントオーバーレイは，特徴的な書式のスライドを検出することでセグメントを抽出する．セグメントを抽出することでトピックの起点が明らかになる．さらに，このセグメントを再帰的に抽出することによって，階層的な構造を求める．これにより，各スライドの節の階層の深さを求める．

3.1 スライドタイプ

提案手法では，スライドを 10 種類に分類している．ここでは，スライドの種類をスライドタイプと呼ぶ．スライドタイプを表 1 に示す．表 1 にはスライドタイプとその略記号，説明，起点を記載している．起点とは，新しいセグメントの起点となるか否かを表し，起点となるスライドタイプの行には「S」，前のスライドと連続して 1 つの内容を説明するスライドタイプの行には「C」を，その他には「-」を記載している．各スライドはいずれかのスライドタイプに分類され．複数のスライドタイプに合致するスライドは，表の上方が優先される．

表 1　スライドタイプの一覧

タイプ名	記号	説明	起点
表紙	FS	表紙のスライド	S
ヘッドライン	HS	実質的な内容がタイトルしかないスライド	S
目次	TS	目次が書かれたスライド	S
節番号	SS	節番号が書かれたスライド	S
目次参照	TRS	目次に記載された見出しをタイトルとするスライド	S
変化点	COS	階層の深さや節が切り替わるスライド	S
リスト	LS	後方の複数のスライドのタイトルを含むスライド	-
リスト参照	LRS	リストに記載された項目をタイトルとするスライド	S
継続	CS	タイトルで前のスライドの続きと判断できるスライド	C
その他	OS	上記以外のスライド	-

3.2 スライドタイプの分類

様々なスライドタイプを含むプレゼン資料の例を図 2 に示す．各スライドがどのスライドタイプになるかは，タイトルaの文字列や位置などを基にルールにより決定する．ルールの詳細は割愛するが，各スライドタイプを検出する基準について概説する．

a　タイトル用のテキストボックスを利用しているか否か，文字の相対的な大きさの違い，テキストボックスの数や位置などを変数としてルールで特定している．

Vol.2010-DD-74 No.6
2010/1/29

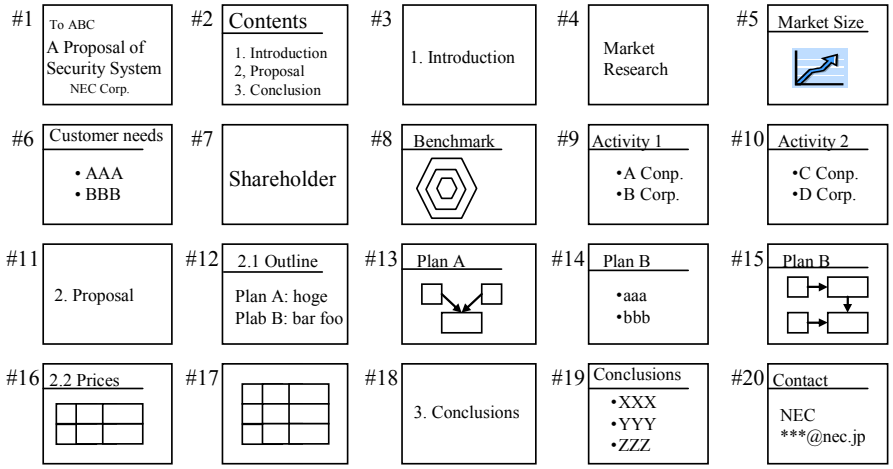

図 2　プレゼンテーション資料の例

FS と HS は大きなトピックの開始を表し，それ以降のタイトルを表示する．HS は，(i)スライド中央にタイトルがある，(ii)表紙用のテンプレートを利用している，などの特徴により検出できる．HS であり，かつ 1 枚目のスライドであれば，そのスライドを FS とする．図 2 の例では，#1 が FS，#3，#4，#7，#11，#18 が HS に相当する．

TS と SS は，目次相当情報であるため推定に役立つ．TS は，「目次」「アウトライン」などの文字列を辞書に登録しておき，これら文字列と一致するスライドを検出する．TS が検出された場合，そのスライド内から目次の見出しを抽出する．TS の後方で，その見出しに合致するタイトルを有するスライドを TRS とする．1 つの見出しに対して対応するタイトルのスライドが複数ある場合は，先頭のもののみを TRS とする．図 2 では，#2 が TS，#3，#11，#18 が TRS に相当する．SS はタイトル中に明示的に節の番号が記載されているものである．第 1 章，1-1，など節番号の表記となるテンプレートとのマッチングにより検出する．#3，#11，#12，#16，#18 などが SS に相当する．

LS と LRS は部分的な階層構造を生成する際に利用され，例えば図 2 の#12～#14 の関係を有するスライドである．LS は，箇条書きや類似する強調を行っている文字列を含むスライドで，その文字列が後方のスライドのタイトルとなっているものを検出する．上記の後方のスライドが LRS となる．

CS は，タイトルに同一性があるスライドを一つにまとめるために検出する．同一性とは，i) 同一のタイトルが連続している，ii) タイトルの後に連番がつけられている，iii)「Continued」「続き」というタイトルである，iv) タイトルの無いスライドが続いている，である．図 2 では，#10，#15，#17 などが CS に相当する．

なお，COS と OS は，アルゴリズム上必要であるが，検出するものではない．

3.3 セグメントの階層化

セグメントオーバーレイでは，下記に示すセグメントの階層化における R1～R5 の 5 つのルールを用いる．スライドタイプの検出時に，以下のルールに従いセグメントを階層化する．下記ルールは，検出するスライドとその他のスライドとの相対的な階層関係を記述するものになっている．

R1: FS は，全スライドの目次構造において根となる．FS は，セグメントにおいて，その他のスライドより 1 つ以上，上位階層になる．

R2: HS は，セグメント内のルートとなる．HS はセグメントにおいて，その他のスライドより 1 つ以上，上位階層になる．

R3: TRS は TS に記載された見出しの間にある階層関係を保持するようにする．

R4: SS は，セグメント内のスライドにおいて節番号の深さb が最も小さいものとの相対的な深さの差により階層化する．

R5: LS は LRS よりも 1 つ以上，上位階層になる．

3.4 アルゴリズム

セグメントオーバーレイのアルゴリズムを以下に示す．

(1) 全てのスライドを一つのセグメントとみなし，全てのスライドの階層の値を 0，スライドタイプを OS とする．

(2) 検出するスライドタイプ（以降，検出タイプ）を設定する．初期の検出タイプは FS である．

(3) 処理対象となるセグメントから検出タイプに合致するスライドを検出する（以降，検出スライド）．検出スライドのスライドタイプを書き換える．
検出できなければ(6)へ．

(4) 従属性のルールを用いて，検出スライドとセグメント内の他のスライドとの階層の深さの差分を求める．階層が深い方が従属するスライドである．

(5) 検出タイプが起点となるものであれば（表 1 の「起点」行が「S」），検出スライドの前方と以降でセグメントを分割する．
このとき得られるセグメントが次の処理対象となる．

(6) 検出タイプを変更．検出タイプがなければ終了．

(7) (2)へ戻る．

b「1.2」なら 2，「2-3-1」なら 3 というように，節番号で階層を表現するための区切り領域の数である．

(2)の検出タイプの初期値は{FS}であり,(5)の処理において{TS と TRS}, {SS}, {HS}, {LS と LRS}, {CS}の順に変更する．この順序は，目次構造を得るための情報の確からしさ，および，形成するセグメントの大きさを勘案し決めている．

図 2 のプレゼン資料に提案手法を適用した際に，検出タイプ毎の処理によって得られる「スライドタイプ（T 列)」,「階層の深さの相対的変化量（R 列)」,「各スライドの階層の深さ（D 列)」を表 2 に示す．T 列において，各処理における検出スライドを太字で記載している．R 列は，検出スライドに対する階層の深さの相対的変化量を記載している．書式は，値が 0 である場合を空欄，値が 0 でない場合を「+数字」としている．D列は，各処理の終了時における各スライドの階層の深さである．また，各処理が終了した時点でのセグメントを太い罫線で示している．提案手法では，各スライドの階層の深さ（D列の値）は，各検出スライドで特定された R 列の値の総和により計算される．この表からも，検出スライドの特定により徐々に目次構造の階層が深まっていく様子を把握できる．

アルゴリズムの処理手順に従って，(3)〜(6)の処理を述べる．

まず，(3)の処理において FS が検出される．なお，検出スライドが FS，HS，TS である場合，次のスライドのスライドタイプは便宜上 COS を与える．次に，(4)の処理により，この時点のセグメント（全スライド）における FS とその他のスライドの階層差「+1」が R の値として記録される．この階層性は 3.3 節の R1 によって規定される．次に，(5)の処理によって，#1 と#2〜20 の 2 つが次のセグメントとなる．#1 はスライド集合が 1 つしかないため処理対象とならない．表 2 において，処理対象から除外されるセグメントはセルを黒く塗りつぶしている．続いて，(6)の処理により検出スライドが TS と TRS に変わり，(2)に戻り同様の処理が繰り返される．検出タイプが HS の処理では，例外処理として，セグメント内に HS が n 回連続して存在する場合，最初の HS 以外は HS とみなさず，検出タイプを HS に固定したまま(3)〜(5)までの処理を n 回繰り返す．これにより，HS を連続して用いてトピックの階層化をしているプレゼン資料の目次構造を推定できる．図 2 の例では，SS の検出が終わった時点でセグメント内に HS が 2 連続する箇所があるため，HS の処理が 2 度行われている．

図 2 のプレゼン資料に対して最終的に推定される目次構造を表 3 にまとめる.表 3 では，各スライドに対する「階層の深さ」,「スライドタイプ」,「起点」を記載している．この階層の深さとトピックの起点のセットが目次構造となる．この目次構造を模式的に表したものを図 3 に示す．図 3 において，一つの枠内に複数のスライド番号が記載されているものは，タイトルが継続しているスライド（表 1 の CS）を意味する．この木構造における部分木がセグメントを表す．図 3 から，提案手法を用いることにより，3.3 節の R1〜R5 を満たす目次構造が推定できていることがわかる．また，スライド#3〜#10 や#11〜#17 に着目すると，目次相当情報がないスライド群についても構造化が行えていることがわかる．

以上のように，セグメントオーバーレイは，HS や LS などの検出タイプを利用し目次構造情報を含まないプレゼン資料を構造化する．また，含まれている検出スライドに基づいて相対的な階層化を順次行うため,ある文字列が含まれているスライドを§1 とするなどといった固定的な目次構造の推定とはならない．このように，提案手法は従来技術の問題を克服できる．

表 2　目次構造の推定過程

	FS			TS+TRS			SS			HS(1 回目)		
#	T	R	D	T	R	D	T	R	D	T	R	D
1	**FS**		0			0			0			0
2	**COS**	+1	1	**TS**		1			1			1
3		+1	1	**TRS**		1	**SS**		1	**HS**		1
4		+1	1			1			1	**COS**	+1	2
5		+1	1			1			1		+1	2
6		+1	1			1			1		+1	2
7		+1	1			1			1		+1	2
8		+1	1			1			1		+1	2
9		+1	1			1			1		+1	2
10		+1	1			1			1		+1	2
11		+1	1	**TRS**		1	**SS**		1			1
12		+1	1			1	**SS**	+1	2	SS		2
13		+1	1			1		+1	2			2
14		+1	1			1		+1	2			2
15		+1	1			1		+1	2			2
16		+1	1			1	**SS**	+1	2	SS		2
17		+1	1			1		+1	2			2
18		+1	1	**TRS**		1	TRS		1	TRS		1
19		+1	1			1			1			1
20		+1	1			1			1			1

Vol.2010-DD-74 No.6
2010/1/29

（表 2 の続き）

	HS(2 回目)			LS+LRS			CS		
#	T	R	D	T	R	D	T	R	D
1			0			0			0
2			1			1			1
3			1			1			1
4	**HS**		2			2			2
5	**COS**	+1	3	COS		3	COS		3
6		+1	3			3			3
7	**HS**		2			2			2
8	**COS**	+1	3	COS		3	COS		3
9		+1	3			3			3
10		+1	3			3	**CS**		3
11			1			1			1
12	SS		2	SS		2			2
13			2	**LRS**	+1	3			3
14			2	**LRS**	+1	3	LRS		3
15			2			2	**CS**	+1	3
16	SS		2	SS		2	SS		2
17			2			2	**CS**		2
18	**HS**		1			1			1
19	**COS**	+1	2	COS		2	COS		2
20		+1	2			2			2

表 3　推定された目次構造の階層の深さと起点

#	階層の深さ	スライドタイプ	起点	#	階層の深さ	スライドタイプ	起点
1	0	FS	S	11	1	SS	S
2	1	TS	S	12	2	SS	-
3	1	HS	S	13	3	LRS	S
4	2	HS	S	14	3	LRS	S
5	3	COS	S	15	3	CS	C
6	3	OS	-	16	2	SS	S
7	2	HS	S	17	2	CS	C
8	3	COS	S	18	1	HS	S
9	3	OS	-	19	2	COS	S
10	3	CS	C	20	2	OS	-

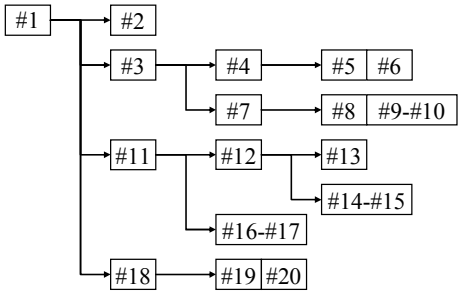

図 3　推定された目次構造

4. セグメントオーバーレイ法の評価

4.1 実験結果

実験に用いるプレゼン資料は，社員 2 万人規模の営業支援共有サイトにアップロードされているプレゼン資料から無作為に抽出した．また，その中で目次が記載されたスライドやタイトルに節番号を含まない，すなわち目次相当情報を含まない 24 件のプレゼン資料（総スライド数 1039 枚）を評価の対象とした．

これらプレゼン資料は，スライドの平均枚数が 43.3 枚と比較的長く，複数のトピックを含む．筆者らは，上記のプレゼン資料の内容を複数人で精査し，トピックやそのサブトピックを抽出していくことで，各スライドの階層の深さを判定し，正解セットを作成した．

評価指標には(1)式を用いた．この式は，各資料において階層の深さが平均的にどの程度正しく求められているかの比率を求めるものである．

$$Fitness(depth) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i} comp\left(depth_i[j],\delta_i[j]\right)\right) \qquad (1)$$

式中の $depth_i[j]$は，目次構造推定によって推定された i 番目のプレゼン資料における j 番目のスライドの階層の深さである．また，$\delta_i[j]$は正解セットの階層の深さを表す．$comp(depth_i[j], \delta_i[j])$は $depth_i[j]$と$\delta_i[j]$の値が等しければ 1，等しくなければ 0 を返す関数である．M はプレゼンテーション資料の数，N_i は各プレゼンテーション資料のスライド枚数を表す．この評価指標は上位階層で一つ階層の深さを間違えるとそれ以下の階層が全て間違いと判定されるため，比較的厳しい指標である．

本評価指標により得られた正答率は 59.8%であった．

4.2 実験結果に対する考察

従来技術では目次構造の推定が困難な目次相当情報のないプレゼン資料に対して，提案手法が 6 割の正答率で目次構造を推定できることを確認した．この正答率は十分とは言えないが，どのようなプレゼン資料からでも自動的に目次構造を推定できる可能性を示すものであると考える．正答率が 6 割程度となった原因の一つは，本手法が正解セットに比べ浅い階層化しか行えていなかったことにある．事実，階層の深さが 0，1，2 であるスライドのみの正答率は 78.5%であり正答率が上がる．

深い階層となるのは，トピックの分類が詳細に行われている状態を表す．しかしながら，提案手法では，スライドタイプの判定を，タイトルの文字列の表層的な特徴に基づいている．これはプレゼン資料においては読者に構成を伝えるために，スライドを代表する文字列のタイトルに，見た目でわかるような工夫をすることが多いと考えたためであるが，この情報だけでは，トピックの細かな分類には至らなかったといえる．例えば，実際のプレゼン資料では，タイトル以外にサブタイトルを記述し，階層性を明記するなど，タイトル以外のテキストボックスを解析する必要があるものも少なからず存在している．また，タイトルの文字列の意味を考慮しないと正しい階層化が行えないものも多い．例えば，1 枚目のタイトルが「**の概要」，それ以降のスライドが「特長」,「仕組み」など**に関する詳細を順次記載する場合，意味的には,「概要」のスライドとそれ以降のスライドは階層構造にあると考えられる．また，タイトルだけではなく，スライドに記載された内容を考慮すれば分類できるトピックもある．人手による正解セット作成は，これら全ての要素を勘案しているため，トピックを細かく分類でき，階層の深いものが含まれている．

このような深い階層を分析するためには，現状の表層情報だけではなく，スライド内の文字列を考慮したスライドタイプの導入やタイトルの文字列の意味解析を行うなどの工夫が必要となる．ただし換言すれば，表層的情報を用いるだけでも，任意のプレゼン資料に対して，3 階層程度（図 1 の§1.1 のレベル）の粒度で目次構造を推定できたことは，提案手法が，プレゼン資料の作成においてタイトルの見た目で概要を分かりやすくするようなノウハウをうまく取り入れられたことを支持するものといえる．

5. 推定した目次構造の検索システムへの適用

提案手法で推定した目次構造の有用性を検証するために，アウトライン・ランキング [8]を実装した社内文書検索システムを開発した．アウトライン・ランキングは，目次構造を利用するランキング技術であり，検索キーワードに適合する節を特定して，その節内に含まれるコンテンツの量により文書の重要度を求めるものである．この手法は，目次構造が正しく求められていると，その効果を発揮する．そこで，このランキングの精度を測ることで推定した目次構造の有用性を間接的に評価する．

本節にてアウトライン・ランキングとそれを用いた検索システムについて述べる．システムの構成を図 4 に示す．本システムでは，通常の全文検索に加え，プレゼン資料をスライド単位で検索するためのスライド単位の全文インデクス（スライド用インデクス）を用いる．また，アウトライン・ランキングのための前処理である目次構造推定とスライド重要度判定，検索時にスコアリングするための適合スライド特定と文書スコア算出の計 4 つのモジュールを用いる．文書検索システムが検索結果を表示するまでの処理手順について述べる．

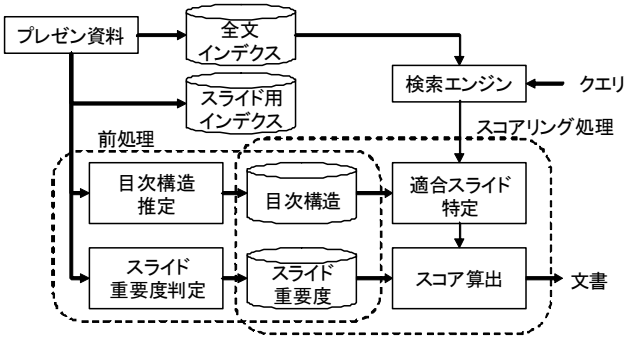

図 4　文書検索システム

5.1 前処理

本検索システムではインデクスを作成する以外に前処理として以下の処理を行う．

- 目次構造推定

各プレゼン資料の目次構造をセグメントオーバーレイにより推定する．

- スライド重要度判定

各スライドに含まれるフローチャートの数，図の数，グラフの数，表の数，文章の数などを求め，これら各コンテンツの重視度を重みとする加重和により，静的なスライド重要度を定量化する．

5.2 スコアリング処理

全文インデクスを利用して，クエリに適合するプレゼン資料を特定する．ヒットした文書をヒット文書と呼ぶ．

- 適合スライド特定

クエリを検索キーワード毎に分ける．次に，スライド用インデクスを用いて，各検索キーワードがヒット文書のどのスライドに含まれるかを特定する．この特定されたスライドに従属するスライドを目次構造により特定し，それらを各検索キーワードに

対する適合スライド集合とする．1 つの検索キーワードが複数のスライドに含まれる場合は，それら全てのスライドに従属するスライドの和集合を検索キーワードに対する適合スライド集合とする．最後にクエリで用いられている検索キーワード間の論理演算を抽出し，各検索キーワードに対する適合スライド集合に対して，抽出した論理演算を適用する．

例えば，図 5 の目次構造において，#1 に SaaS，#2，#3 に市場規模という文字列が含まれているとする．このとき，クエリ「SaaS AND 市場規模」で検索されたとする．上記アルゴリズムに従えば，SaaS に対する適合スライド集合は全スライド，市場規模に対する適合スライド集合は#2，#3，#4 となる．両検索キーワードは AND 演算で結ばれているため，各々の適合スライド集合にも AND 演算を適用し，最終的に適合スライド集合#2，#3，#4 を特定する．図 5 に適合スライド集合の特定結果を示す．このようにアウトライン・ランキングでは，目次構造に基づいて，クエリに対する適合スライド集合を決定する．適合スライド集合は，スコアの算出に用いられるため，結果としてアウトライン・ランキングの精度は目次構造の正しさに依存する．

- スコア算出

クエリに対するプレゼン資料のスコアを適合スライド集合のスライド重要度の和とする．図 5 にスコアの算出イメージを示す．図の各スライドの右横に書かれた数字はスライド重要度を表す．本例では，#2，#3，#4 の静的なスライド重要度がそれぞれ 5，2，1 であるため文書スコアは 8 となる．なお，本システムでは，目次構造上においてより上位階層に検索キーワードが含まれているとスコアを高くするなどの補正を行っている．上記の処理を全てのヒット文書に行い，スコアの高い順に提示する．

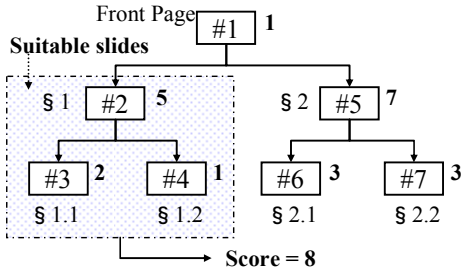

図 5　スコア算出

5.3 ユーザインターフェース

アウトライン・ランキングを実装した検索システムの GUI は，一般的な検索システムと同様，クエリを入力し，検索ボタンを押下すると検索結果が表示されるものである．検索結果の画面には，検索キーワードを含むプレゼン資料がスコアの高い順に表示され，各プレゼン資料はファイル名とスニペットが表示される．また，各資料のファイル名の横には，目次表示ボタンが付与されており，検索者は，この目次表示ボタンを押下することで，セグメントオーバーレイにより推定された目次構造を閲覧できる．目次構造の表示画面では，各スライドのタイトルが階層の深さでインデントされて表示され，このプレゼン資料をダウンロードするためのボタンが付与されている．

6. 検索の評価

上記の検索システムを約 2 ヶ月間社内で運用し，その検索ログを用いて，推定した目次構造が検索の効率化に役立つものであることを確認した．検索対象は企業内に蓄積されていた約 17,000 件の PowerPoint ファイルである．以下に，その詳細を述べる．

6.1 実証実験

アウトライン・ランキングは，検索キーワード間の関連性が高いスライドのみを絞り込み，文書の重要度を求める．そのため，検索キーワードが複数であると，それら複数の検索キーワード全てに適合するスライドのみで文書が評価されることになり，単一の場合と比較して，より検索者が必要とする情報が含まれやすい文書を上位にランクしやすくなる．すなわち，検索キーワードが複数である場合の方が，アウトライン・ランキングの効果が発揮されやすい．無論，この効果は，目次構造が正しく得られているこが前提となる．そこで，ユーザが必要とするプレゼン資料がアウトライン・ランキングと従来の TF/IDF ベースによるランキング（以降，従来ランキング）で，それぞれ何位になるかを調べ，その結果を，検索キーワードが単一である場合と複数である場合とでそれぞれ比較した．

6.2 ログデータの収集

できる限り検索者が必要としてアクセスしたプレゼン資料に関する検索ログで目次構造推定の有用性を検証するために，(i)ヒット数が 20 件以上であり，(ii)目次構造の表示画面からプレゼン資料にアクセスした検索ログを収集した．(i)は，検索件数が多いほど，検索者が必要とする資料が含まれる可能性が高いためである．(ii)は，目次構造の表示画面からプレゼン資料にアクセスしている場合，概要を確認したうえでそのプレゼン資料を必要と判断したと考えられるため，通常の検索結果の画面からのアクセスよりも，検索者が欲したプレゼン資料である可能性が高くなる．検索結果画面から直接プレゼン資料にアクセスしている場合，検索者は内容を確認するためにアクセスしていることも多く，アクセスされたプレゼン資料が検索者の欲する情報とは異なることも多い．

プレゼン資料にアクセスがあった検索の中で，(i)(ii)の条件を満たすものを有効検索と呼ぶ．検索ログとして，有効検索のときの検索キーワード，アクセスされたプレゼン資料の表示順位，そのときの検索のヒット数を記録した．

6.3 ログデータの解析

有効検索の回数は 31 回であった．有効検索においてアクセスされたプレゼン資料が，従来ランキングを用いた場合に何位にランクされるかを調べ，検索キーワードが単一／複数の場合のそれぞれで，アウトライン・ランキングによる順位と比較した．なお，有効検索毎にヒットする文書数が異なるため，順位の比較は，ヒット数で順位を割り正規化した値で行っている．正規化された順位は(0:1]の値をとり，アクセスされたプレゼン資料が上位にランクされているほど 0 に近づく．

31 件の有効検索のうち，単一のキーワードによる検索数が 20 件，複数のキーワードによる検索数が 11 件であった. 単一／複数の検索キーワードにおける正規化された順位の比較結果を，それぞれ表 4 と表 5 にまとめる．平均値に対して，片側 t 検定により, その差を比較したところ, 単一／複数の検索キーワードで, それぞれ p 値が 0.154, 0.001 であった．

表 4　単一キーワード時の正規化された順位の比較

	アウトライン・ランキング	従来ランキング
平均（分散）	0.13 (0.03)	0.21 (0.09)

表 5　複数キーワード時の正規化された順位の比較

	アウトライン・ランキング	従来ランキング
平均（分散）	0.17 (0.051)	0.39 (0.124)

6.4 考察

表 4 と表 5 に示す平均値の値から，検索キーワードが単一／複数どちらにおいても，アウトライン・ランキングの方が有効検索においてアクセスした資料を平均的に高い順位でランクしている．ただし，本検索システムでは，アウトライン・ランキングで検索結果を表示しているため，検索結果の上位のものしかアクセスしないなどの検索の癖を考慮すると，アウトライン・ランキングと従来ランキングにおける順位の差の比較はフェアとは言い難い．

ただし，検索キーワードが単一である場合と複数である場合において，上記の検索の癖による順位の差は変化しないはずである．しかしながら，単一キーワードと比較すると，複数キーワードの検索においては p 値の値が 2 桁小さくなっており，その差が有意になっている．すなわち，複数の検索キーワードであると，アウトライン・ランキングが従来ランキングよりもより高順位で必要な文書を提示しやすくなる．これは，複数のキーワードを用いることによって，アウトライン・ランキングによる適合スライド集合の絞込みの効果が強く現れ，検索者が必要とする内容を含む文書を正しく評価できたためと考えられる．適合スライド集合の絞込みは，プレゼン資料の目次構造が正しく求められていることで実現できる．そのため，上記の結果は，セグメントオーバーレイにより推定した目次構造が，アウトライン・ランキングを用いた検索の効率化に寄与することを示唆するものである．

7. まとめ

プレゼン資料の目次構造を推定するセグメントオーバーレイを提案した．実験により，従来技術では目次構造を抽出できなかったプレゼン資料から約 60%の正答率で目次構造を推定できることを示した．また，提案手法で推定した目次構造の有用性を検証するために，アウトライン・ランキングを実装した社内文書検索システムを構築し実証実験を行った．検索ログを解析し，提案手法により推定した目次構造が，検索の効率化に寄与するものであるとの示唆を得た．

目次構造の推定技術を進展させれば，資料の作成能力の客観的な評価にも利用可能であると考えている．目次構造が推定しやすいプレゼン資料は，構成が確かな文書であると考えられる．そのため，目次構造推定技術は，資料の構成の良し悪しや，一般的な構成とのズレを定量化できる．このような評価を可能にし，作成者に資料の構成を見直す機会を与えられれば，蓄積されたプレゼン資料を，資料の作成能力の向上にも役立てられることが期待できる．今後は，目次構造技術の精度の向上を図るとともに，上述のような技術の水平展開を行っていきたい．

参考文献

1) M. A. Hearst, TextTiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics, Vol. 23, No.1 pp.33-64, 1997
2) 平尾，北内，木谷，語彙的結束性と単語重要度に基づくテキストセグメンテーション，情報処理学会論文誌．データベース, 41(SIG_3(TOD_6)), pp.24-36，2000
3) Jeffrey C. Reynar. Statistical Models for Topic Segmentation. In Proc. of ACL-99, pages 357–364. 1999
4) D. M. Blei and P. J. Moreno. Topic Segmentation with an Aspect Hidden Markov Model, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp.343-348, 2001
5) G. Dias and E. Alves, Unsupervised Topic Segmentation Based on Word Cooccurrence and Multi-Word Units for Text Summarization, In Proc. of the ELECTRA Workshop associated to 28th ACM SIGIR Conference, pp.41-48, 2005
6) 三木，教材スライド間の類似性に基づく講義の構造分析，京都大学 特別研究報告書，2003
7) S. Klink, A. Dengel, and T. Kieninger, "Document Structure Analysis Based on Layout and Textual Features," in DAS - International Conference of Document Analysis Systems, pp. 99-111, 2000
8) 山本，松田，“社内文書検索システム（４） -セグメントオーバーレイによるプレゼンテーション資料からの目次構造特定-”, 第 70 回情報処理学会全国大会, 2008