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-

山本 康高†, 松田 勝志†
NEC サービスプラットフォーム研究所†

1 はじめに

近年，企業では Microsoft PowerPoint などで作成できるプレゼンテーション資料(以降，プレゼン資料)が増加しており，それらプレゼン資料の活用が社内知の有効活用のために重要になってきている．その方法の一つに，目次構造を特定しておくことが挙げられる．目次構造とはプレゼン資料の明示的/非明示的な節の構成である．目次構造は，資料の節単位の検索，検索結果において目次をスニペットとして表示，節毎の保存管理，など様々な用途に利用できる．しかしながら，従来の目次構造特定技術は，以下の課題があり目次構造を特定できるプレゼン資料が限られている．

- 目次やタイトル中の節番号「1. ○○」などの目次に相当する情報(目次相当情報)が明記されていないと目次構造を特定できない．
- ある文字列が含まれているスライドを§1 とみなすなど，スライドの特徴と節とが 1 対 1 対応でしか目次構造を特定できない．

一方，目次構造特定に類似する技術として単語の共起性や頻度などに基づいて文書の節の切り替わりを検出する手法がある[1][2][3]．しかしながら，これらは，テキストを分割する手法であり，節内にさらに節を含むような節の包含関係まで特定することはできない．ここで節をあるトピックに関するスライドの集合と定義しセグメントと呼び，セグメントの包含関係を従属性と呼ぶ．

本稿では，任意のプレゼン資料の目次構造を特定できるセグメントオーバーレイを提案する．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しセグメントと従属性の特定を再帰的に行い，目次構造を特定する．実験により，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特定できることを示す．

2 目次構造

プレゼン資料の全体を代表するスライドを表紙と捉えると，プレゼン資料は図 1に示す階層構造により表現できる．「#数字」はスライドのページ数，「§数字」は節の番号を表す．

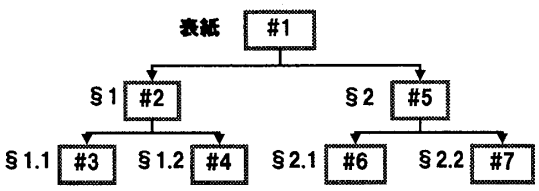

図1 目次構造

本稿では，表紙をルート(根)とする木構造として目次構造を表現する．#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．ただし，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特定するための情報源になり得る．ノウハウの例を示す．

- 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することで，暗黙的にセグメントを形成する．
- 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，暗黙的に従属性を形成する．

明記された目次相当情報に加え，これら特徴的な書式を検出すれば詳細な目次構造が特定できる．

3 セグメントオーバーレイ

3.1 スライドタイプ

セグメントオーバーレイではスライドの種類を書式の特徴で分類し，スライドの種類に応じてセグメントと従属性を特定する．スライドの種類をスライドタイプと呼ぶ．スライドタイプの一部を表1に示す．表には8種類のスライドタイプと略記号と説明文を記載している．また，セグメントの起点となるスライドタイプの「起点」行には○を付与している．

表1 スライドタイプ

スライドタイプ	略記号	説明	起点
表紙スライド	FS	表紙のスライド	○
ヘッドライン スライド	HS	実質的な内容が タイトルしかないスライド	○
目次スライド	TS	目次が書かれたスライド	○
節番号スライド	SS	節番号が書かれたスライド	○
目次参照 スライド	TRS	目次に記載された見出しを タイトルとするスライド	○
リストスライド	LS	そのスライド以降の複数の スライドのタイトルを含むスライド	
リスト参照 スライド	LRS	リストスライドに記載された項目を タイトルとするスライド	○
連続スライド	CS	タイトルから前のスライドの続き と判断できるスライド	

An In-house Documents Retrieval System (4) -Identifying Table of Contents Structure of Presentation Materials using Segment Overlay-
†Kosuke YAMAMOTO and Katsushi MATSUDA,
Service Platforms Research Laboratories, NEC Corporation

3.2 従属性のルール

目次構造を特定する際には，どのスライドタイプがどのスライドタイプを従属するかを決めておく．従属性のルールの一例を以下に示す．

R1:表紙スライドは目次構造においてルート(根).
R2: リスト参照スライドはリストスライドに従属.

3.3 手法の概要

セグメントオーバーレイの概要を以下に示す．なお，各スライドがどのスライドタイプに合致するかは，各スライドのタイトルやタイトルの位置などで決定する．

(1) 全てのスライドを一つのセグメントとみなす．検出するスライドタイプを設定にする．これを検出タイプと呼ぶ．初期値はFSである．
(2) 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．これを検出スライドと呼ぶ．検出できなければ(4)へ．
(3) 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，従属性のルールに基づいて求める．階層が深い方が従属されるスライドである．
(4) 検出タイプが起点となるものであれば(表1の起点行が「○」)，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．
(5) 検出タイプを変更する．
検出タイプがなければ終了.
(6) (2)へ戻る.

検出タイプは，初期値をFSとし，(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．最初は，表紙スライドとその他のスライドの従属性を特定する．セグメントの従属性は，階層の深さにより表現する．例えば，7枚のスライドからなるプレゼン資料を想定する．そのプレゼン資料で#1が検出スライドとなった場合，(3)の処理では，スライド#1の階層の深さを「0」とし，その他のスライド#2～#7の階層の深さを「1」とする．これは，スライドタイプ「FS」以外のスライドは「FS」のスライドに比べ，少なくとも1つ階層が深いことを表す．この従属性は3.2節のR1によって規定される．また，階層の深さが「1」である#2～#7のスライドが次の処理対象となるセグメントになる．同様の処理を，検出タイプを変えて行い，その都度階層化を行いながらセグメントを絞り込み，目次構造を特定する．

検出タイプにはHSやCSなど明示された目次ではない情報も含まれる．また，検出スライドがあればそこから相対的に階層化されるため，ある文字列が含まれているスライドを§1とするなどの固定的な特定とはならない．このように，セグメントオーバーレイは従来技術の課題を克服している．

なお，従来手法の多くは，本手法におけるTSとTRSのみ，もしくはSSのみを用いた目次構造特定であるといえる．このように単一のスライドタイプのみで目次構造を特定している理由は，各スライドタイプで得られる目次構造が矛盾した場合への対処が困難であるためと考えられる．しかしながら，本手法は，各スライドタイプによる構造化を，特定されているセグメント内に限定し順次行っていくことで，この課題を解決している．これも本手法の特徴の一つである．

4 評価

目次相当情報のない 24 件のプレゼン資料に対して，筆者らが各スライドの階層の深さを判断し，正解セットを作成した．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を評価指標とした．これ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，比較的厳しい評価指標である．

実験の結果，正答率は 59.8%であった．従来技術では目次相当情報のないプレゼン資料から目次構造を特定することはできない．本実験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特定できるという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．正解セットの階層の深さが 0，1，2 であるスライドのみを用いた正答率は 78.5%であった．3 階層程度の目次構造(図 1の§1.1のレベル)を特定するという目的ならば本手法は有効であるといえる．これ以上の深い階層化を行うためには，他のスライドタイプを取り入れる，スライド内の文字列の意味解析を行うなどの工夫が必要である．

5 まとめと今後の課題

本稿では，任意のプレゼン資料から目次構造を特定するセグメントオーバーレイを提案した．本手法では，特徴的な書式のスライドを抽出しセグメントを再帰的に細分化することで目次構造を特定できることを述べた．実験により，約 60%の正答率で目次構造を特定できることを示した．今後の課題として，更なる精度向上，目次構造特定技術の他技術への応用，従来技術との比較などが挙げられる．

参考文献

[1] Hearst. M. TextTiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics, Vol. 23, No.1 pp.33-64, 1997
[2] 戸田, 北川, 藤村, 片岡, 奥, “グラフ分析を利用した文書集合からの話題構造マイニング” 電子情報通信学会論文誌, Vol.J90-D, No.2, pp.292-310, 2007
[3] 三木 健司，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