社内文書検索システム（５）
-プレゼンテーション資料の目次構造を利用したアウトライン・ランキングの提案-

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

企業における社内文書検索の目的の一つは，他の文書の一部を再利用することで資料作成を効率化することにある．IDC の調査[1]からも資料の再利用に対する時間削減が企業活動の効率化に有効であることがわかる．特に，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)は，図表やグラフなど価値あるコンテンツを多く含む．ただし，プレゼン資料内に多くのコンテンツが含まれているからといって，その資料が再利用しやすいわけではない．なぜならば，検索者の欲する情報はプレゼン資料の全体ではなく，資料の一部のスライドに書かれていることが多いためである．すなわち，プレゼン資料における再利用性とは，検索クエリに適合するスライドにコンテンツが豊富に含まれていることと定義できる．

再利用性という観点をランキングに反映させた文書検索システムは見当たらない．一方，検索キーワードの文書中での位置や検索キーワード間の文法的関係を考慮し，ランキング精度を向上させる手法は報告されている[2][3]．しかしながら，これらの手法では，プレゼン資料内において検索クエリに適合する内容のスライドが資料中にどの程度含まれているかを特定することはできない．

本稿では，再利用性の高いプレゼン資料を上位にランクさせるアウトライン・ランキングについて述べる．また，簡易実験によりコンテンツが豊富な再利用性の高い文書を上位にランクできることを示す．

2 アウトライン・ランキングの概要

目次構造の例を図 1に示す．図 1は 7 枚のスライドからなるプレゼン資料であり，「#数字」はスライドのページ数，「§数字」は節の番号を表す．

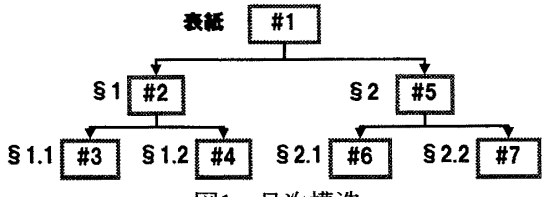

図1 目次構造

図 1では表紙をルートとし，各節の従属関係が木構造として表現されている．ある検索者が検索クエリ「NGN AND 市場規模」で検索したことを想定する．図 1において，表紙である#1 のタイトルが「NGN」，§1 である#2 のタイトルが「市場規模」であるとする．このとき，§1 は「NGN の市場規模」に適合するものと判断でき，同様に§1 より下位のスライド(#3，#4)も，この検索クエリに適合するものと捉える．このときスライド#3，#4 は NGN や市場規模という文字列を含まなくて良い．これは，§1 に従属するスライドは§1 の内容を受けた詳細であるためである．すなわち，あるスライドの文字列は，目次構造におけるそれ以下のスライドにも暗黙的に含まれる文字列と考える．目次構造を用いることで，各スライド内の検索キーワードの有無に関係なく，検索クエリに適合するスライドの集合(以降，適合スライド集合)を特定できる．

この適合スライド集合のスライド重要度の和によって検索クエリに対する文書の重要度(スコア)を算出する．スライド重要度とは，スライドに含まれる図表やグラフなどのコンテンツの量に基づいて定量化する各スライドの重要度である．

3 アウトライン・ランキングによる文書検索

図 2にアウトライン・ランキングを用いた検索システムの全体像を示す．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，検索時にスコアを算出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．

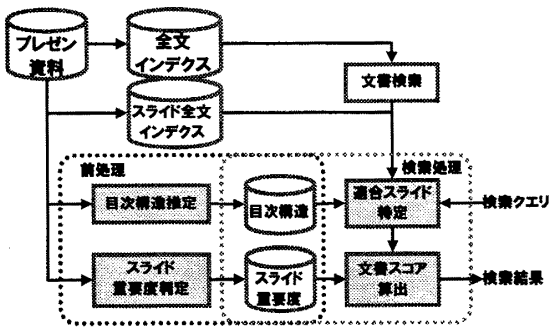

図2 文書検索システム

An In-house Documents Retrieval System (5) -Proposal of Outline Ranking using Table of Contents Structure of Presentation Materials-
†Kosuke YAMAMOTO and Katsushi MATSUDA,
Service Platforms Research Laboratories, NEC Corporation

3.1 前処理

本文書検索システムでは，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求める．目次構造はセグメントオーバーレイ[4]により推定する．詳細は文献[4]を参照されたい．

スライド重要度判定：各スライドから，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，各コンテンツの重視度を重みとする加重和によりスライド重要度を定量化する．

3.2 検索処理

全文インデクスを利用して，検索クエリに適合するプレゼン資料(ヒット文書)を特定する．

適合スライド特定：検索クエリを検索キーワード毎に分ける．次に，スライド全文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．一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．最後に検索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 1の目次構造において，#1 に NGN，#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，検索クエリ「NGN 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，NGN 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．各検索キーワード間の論理演算である AND 各検索キーワードの適合スライド集合に適用し，適合スライド集合#2，#3，#4 を特定する．図 3に適合スライド集合の特定結果を示す．

文書スコア算出：適合スライド集合のスライド重要度の和を求める．図 3にスコアの算出イメージを示す．図 3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．この処理を全てのヒット文書に行い，スコアの高い順に検索者に提示する．

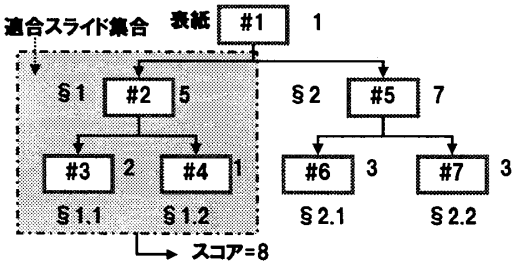

図3　スコア算出

提示する際は適合スライド集合にダイレクトにアクセスできるようにすることもできる．これにより，求める情報へのピンポイント検索を支援することができる．これはアウトライン・ランキングならではの効果の一つである．

4 評価

アウトライン・ランキングを実装した文書検索システムを開発した．この文書検索システムに約 400 件のプレゼン資料を登録し簡易実験を行った．検索クエリ「NGN AND 市場規模」に対して最上位にランクされた文書の適合スライド集合の 1 枚を図 4左に示す．図中の破線で囲まれている部分を拡大したのが図 4右であり，これは NGN の市場規模に関するグラフである．このスライドは NGN という文字列は含まれていないため単にスライド単位の文書検索ではこのスライドは発見できない．これは目次構造を用いた適合スライド特定がうまく働いた例といえる．なお，TF*IDF によるランキングではこの文書は最上位にランクされなかった．

実験例は少ないが他の実験でも類似する結果を得ている．この結果は本手法が再利用性の高い資料の検索を支援できることを示唆するものである．

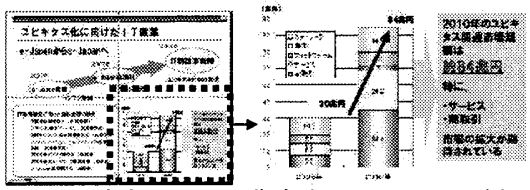

図4　適合スライド集合中のスライドの例

5 まとめと今後の課題

目次構造を利用することで検索クエリに適合するスライド集合を特定し，そのスライド集合内のコンテンツ量で文書の重要度を算出するアウトライン・ランキングについて述べた．簡易実験により再利用性の高い文書を上位にランク可能であることを確認した．ユーザ評価の実施，プレゼン資料以外の文書への本手法の適用などが今後の課題である．

参考文献

[1] 検索コンテンツ管理技術の投資価値に関する調査の調査結果, IDC 2004, http://www.computerworld.jp/news/ sw/48960-3.html

[2] 松本, 小西, 高木, 小山, 三宅, 伊東, “文構造における検索キーワード間の修飾: 被修飾関係に基づく WWW 検索精度の向上” 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション Vol.105, No.595, pp.7-12, 2006

[3] 田, 手塚, 小山, 田島, 田中, “質問キーワードの近接性と密度分布に基づくウェブ検索の改善手法” 日本データベース学会 Letters Vol.5, No.1, pp.113-116, 2006

[4] 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