スライドの流用性に着目した企業内スライド検索手法の提案

有熊　威†　　白石　展久†

ＮＥＣ　サービスプラットフォーム研究所†

1 はじめに

企業内のオフィス文書作成では，既存の文書の一部を有効に再利用して文書作成時間を削減することにより，知的生産活動の生産性を向上させたい要求がある．特にプレゼンテーション資料の改版や概要説明資料の作成では，他者の作成済みスライド群から，再利用可能なスライドを作成資料に取り込み，重複した内容のスライドを作成する無駄を削減したい需要がある．

しかし，現在主流の検索システムは，キーワード出現頻度や文書間の参照情報(例えばウェブにおけるリンク)に基づいており，上記需要を満たすような高再利用スライド(製品や技術の説明に使われる典型的なスライドなど)を検索することは困難であった．

このような背景から，本研究ではプレゼンテーション資料間でのスライドの流用性に基づき，高再利用スライドの検索手法を提案する．

2 スライド流用性に基づく高再利用スライド検索手法

2.1 スライド流用性

高再利用スライドを検索するためには，スライド流用性(そのスライドの流用のしやすさ，少ない修正で作成資料に取り込めるかどうか)をどのように計算機に判断させるかが課題となる．

この課題を解決するために，「流用性の高いスライドほど，他の資料へ流用されることが多い」と仮定し，スライドの流用回数からスライド流用性を推定する．スライドの流用回数は，対象となる既存資料群内で流用されたスライドの集合(流用スライド集合)のスライド数として求める．流用スライド集合は，スライド内容間の類似度が一定の閾値以上のスライドを流用されたスライドとみなして求める(図 1)．

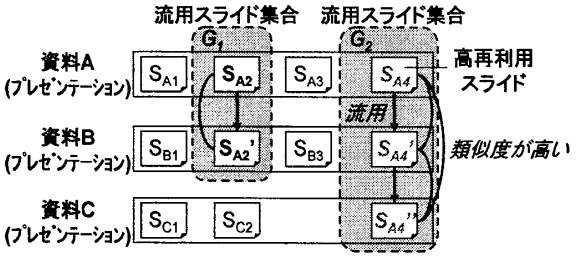

図 1 資料群内の高再利用スライドと流用

Quotablity Oriented Presentation Slides Retrieval Method
† Takeshi Arikuma and Nobuhisa Shiraishi,
Service Platform Research Laboratories, NEC Corporation.

2.2 高再利用スライドの検索手法

スライド流用性を推定し，検索システムに適用することで，検索結果のスライド流用性に基づいたランキングが実現できる．利用者は資料作成時に，作成対象の分野に関連したキーワードで検索することで，高再利用スライドを容易に見つけられるようになる．

高再利用スライドの検索へスライド流用性を適用するために，図 2 のように，検索時に各流用スライド集合について，利用者の検索キーワードに関連したスライドのみの部分集合($G_i' \subset G_i$)を特定し，部分集合(G_i')における流用性を算出する．これは，検索の全対象資料を対象とした流用性を用いると，プレゼンテーションの区切りを示すスライド(「デモ」や「ご参考」など)のように流用する価値の低いスライドの流用性が高く算出されてしまうためである．

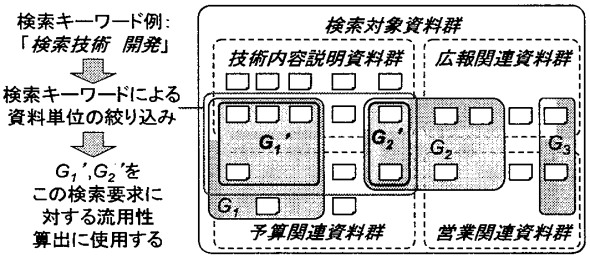

図 2 流用性算出対象スライドの絞り込み例

3 スライド流用性の算出アルゴリズム

スライド流用性の値を算出するために必要な，類似度判定による流用スライド集合の生成(事前に実施)と，流用スライド集合を用いた検索結果資料群でのスライド流用性の算出(検索時に実施)のためのアルゴリズムを提案する．

3.1 流用スライド集合の生成

検索対象資料群から，各スライド間の類似度判定と，流用スライド集合の生成を行う．

1. 各資料をスライド s 単位へ分割し，スライド集合 S を作成
2. スライド $s \in S$ の特徴情報 Ts を抽出
3. スライド $s \in S$ と，$s' \in S$ $(s' \neq s)$との類似度を類似度関数 $\mathrm{sim}(Ts, Ts')$で算出
4. 類似度が閾値 ξ より高いスライドを纏め，流用スライド集合 $G_i(1 \leqq i \leqq$集合数 $n_g)$を作成

特徴情報として，スライド内のテキスト情報を，類似度関数として Levenstein 距離を用いる．類似度判定には，図やグラフなども利用できるが，スライドの要点

はテキストで表現される場合が多いこと，図の類似度判定はテキストより計算量が多ことからテキスト情報を用いる．

3.2 検索結果資料群でのスライド流用性の算出

利用者からの検索要求に対応した検索結果資料群を求め，流用性スコアを算出する．

1. 検索要求 r と適合度の高いプレゼンテーション資料群のスライド集合 P_rを求める
2. 検索要求 r に対するスライド集合 G_iの流用性 q_{Gir}を下記の式で算出する

$$q_{Gir} = \frac{\mathrm{n}(G_i \cap P_r)}{\max_{1 \le j \le n_g} \left(\mathrm{n}(G_j \cap P_r)\right)}$$

3. 流用スライド集合 G_i 内のスライド $s \in G_i$ の流用性スコアを $q_s = q_{Gir}$ として求める

P_r を求める方法としては利用者が入力した検索キーワードによる資料検索を用いる．

4 評価実験

4.1 評価手法

評価実験として，流用性をランキングに使用した場合(提案手法)の実験を実施した．情報の新鮮さを考慮するため，資料の日付情報をスコア化し，流用性スコアに加重加算してランキングを求めた．対照実験として，tf*idf[1]を使用した場合(従来手法)で実験を実施した．従来手法では，各資料のテキストにおける検索キーワード出現頻度の高いスライドが，上位にランキングされることになる．提案手法と従来手法それぞれについて，同じ検索キーワードによる検索を実施し，検索結果のランキングを比較した．

検索結果ランキングの評価指標として，上位 K 件における Normalized Discount Cumulative Gain (NDCG) [2]を用いた．評価用データとして，社内のプレゼンテーション資料約 3,200 件(約 92,000 スライド)を用いた．評価用データセットとして，キーワードと適合度ラベルつきの結果スライド群の組を用意した．適合度ラベルは，0～4 の整数値(Bad, Good, Excellent, Perfect)を用いた．データセットに無い検索結果は適合度が 0(Bad)として NDCG を計算した．

4.2 評価結果と考察

実験結果を図 3 に示す．提案手法(流用性＋日付スコア)の NDCG(K=1～20)は 1 に近い値となり，提案手法が高再利用スライドを上位にランキングすることを確認した．また，従来手法(tf*idf＋日付スコア)と比べ，NDCG(K=10)が約 1.9 倍に改善しており，提案手法は同じ検索キーワードによる検索で，高再利用スライドを従来手法より上位にランキングできることを確認した．

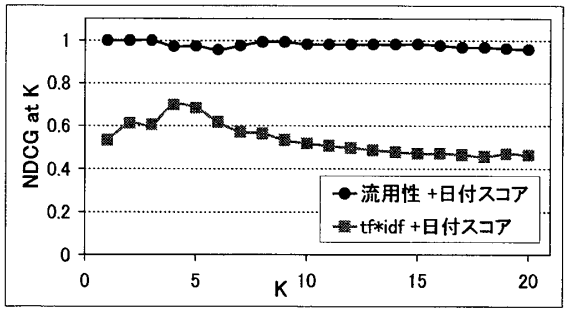

図 3 結果：上位 K 件における NDCG

5 実用性検証システムの試作

本研究で提案・評価した検索手法の実用性を検証するためのシステムを，検索プラットフォーム CRISP[3]を用いて試作している(図 4)．

今後，拡販・商談用資料 DB に提案手法を適用し，検索システムの利用傾向の分析や，社内利用者へのアンケートを通して実用性を検証する．

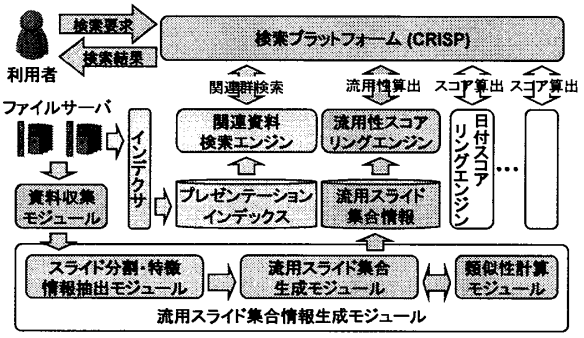

図 4 高再利用スライド検索システム

6 まとめと今後の展望

本研究ではプレゼンテーション資料におけるスライド流用性から，高再利用スライドを検索する手法を提案した．社内プレゼンテーション資料約 3,200 件を対象として評価実験を実施し，検索結果上位 10 件の NDCG が従来手法に比べ約 1.9 倍に改善することを確認した．提案手法により，製品や概念の説明に使われる典型的なスライドの検索が可能になることで，資料作成などの業務効率を改善できると考えている．

今後は，実用性検証システムの実利用を通じて業務効率改善の効果の評価に取り組むと共に，プレゼンテーション資料作成ソフトウェアと連携し，作成中資料に最適な高再利用スライドをダイナミックに推薦する機能などによって，企業内文書の再利用を促進し，更なる業務効率化を実現できると考えている．

参考文献

[1] G. Salton and C. Buckley: Term-weighting approaches in automatic text retrieval, Information Processing & Management, Volume 24, Issue 5, Pages 513-523 (1988).

[2] K. Järvelin and J. Kekäläinen: Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems, Vol.20, No.4, pp.422-466 (2002).

[3] 白石 展久：社内文書検索システム(1)-検索プラットフォーム CRISP-，情報処理学会第 70 回全国大会，pp.1-445-446 (2008).