3D1-R-13-2

プロジェクト提案のための文書情報管理システムの開発

Development of Document and Information Management System for Project Proposal

稗方 和夫[*1] 大和 裕幸[*1] 笈田 佳彰[*1] 岡田 伊策[*2] 齋藤 稔[*2]
Kazuo HIEKATA, Hiroyuki YAMATO, Yoshiaki OIDA, Isaac OKADA, Minoru SAITO

[*1] 東京大学大学院新領域創成科学研究科
[*1] Graduate School of Frontier Sciences, THE UNIVERSITY OF TOKYO
[*2] 富士通株式会社共通技術本部ナレッジ推進統括部
[*2] System Engineering Knowledge Improvement div., SYSTEM ENGINEERING TECHNOLOGY UNIT, FUJITSU LIMITED.

Presentation creation support system was developed. Presentations are managed as a slide unit using URI. Diverse information is attached to each slide by RDF to improve search efficiency. Especially, connecting similar slides based on the slide image and the text in the slide is effective to search enough candidate slides for reuse. Case study illustrates that the time required for presentation creation is reduced by around 20% using the system and the created presentation includes the more various slides which are included in different existing presentation files.

1. 緒言

企業内において、プレゼンテーションは最も重要な業務の一つである。プロジェクトの提案やシステムの説明を行う上で必須な文書である。プレゼンテーション作成業務の中で、既存プレゼンテーションの再利用は欠かせないプロセスであるが、多くのスライドを含む PowerPoint ファイルであるため、目的のスライドを検索し、再利用する際に、プレゼンテーションファイルの開閉や、ファイル内における無関係なスライドの閲覧といった無駄な作業が伴うため、限られた選択肢から再利用するスライドを決定せざるを得ない。多様な課題・要望に応じて柔軟にプレゼンテーションを作成するには困難が伴う。

そこで本研究では、プレゼンテーションファイルをスライド単位に分割し管理の一元化を図り、メタデータを用いて各スライドを有効に結びつけることで既存プレゼンテーションの再利用効率の向上を目指した作成支援システムを開発する。特に、再利用する候補スライドを網羅的に収集し、多様なプレゼンテーションの作成を支援するため、メタデータによる類似スライドを関連づけに主眼を置く。また、実務経験者による利用を通じて開発したシステムの有効性を評価する。

2. プレゼンテーション作成システム

2.1 システム概要

クライアントサーバ型で構築した提案システムの概要図を図 1 に示す。クライアント側からサーバの機能を利用するための各インタフェースについて説明する。

2.2 知識蓄積インタフェース

プレゼンテーションを再利用しやすい形式に変換し、本システムに蓄積するためのインタフェースである。

(1) 文書情報蓄積粒度

本システムではスライド単位とプレゼンテーション単位の 2 つの粒度でプレゼンテーション文書情報を管理する。蓄積粒度の管理には、プレゼンテーションをアップロードする段階で、プレゼンテーションとプレゼンテーションが含むスライド全てに固有の識別子である URI を割り当てる。

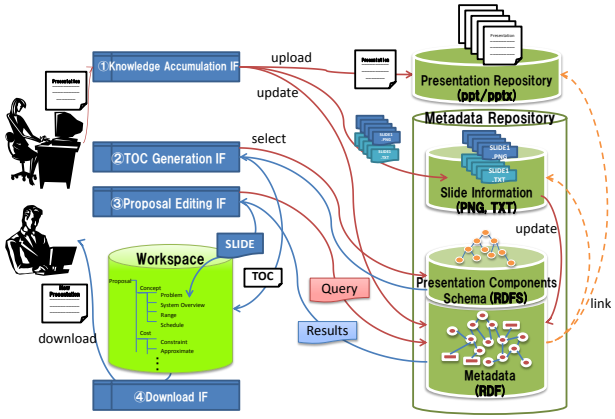

図1 システム概要図

(2) メタデータ付与の概要

URI を割り当てられた文書情報に対して、RDF を用いてメタデータを付与することで属性の記述や、文書情報間の関連付けを行う[神崎 2005] 。以下でメタデータの種類別に説明を行う。

基本情報に関するメタデータはアップロード時に付与され、類似スライドに関するメタデータは、負荷の大きさから定期的なバッチ処理により付与される。

(3) 基本情報に関するメタデータ

アップロード時刻や、ファイルサイズといった基本属性情報に加え、プレゼンテーションにおけるスライドの前後のつながりを表すメタデータを付与する。

(4) 類似スライドの関連付け

対象とする文書情報がスライドであり、スライド自体は画像と見なすことができ、文字列も多分に含む。そのため、類似度を計算する特徴量として、スライド画像由来の形状[Sivic 2003] 、色情報[Swain 1991]に加え、テキスト情報[Eishbein 2008]も用いて類似度を計算し、閾値を超える類似度を持つスライドを RDF により紐付ける。

2.3 目次生成インタフェース

本インタフェースはプレゼンテーション作成において、新規プレゼンテーションの基本構成を決定するために一度だけ使用される。事前にベテラン作成者がプレゼンテーション目次項目ス

キーマから目的別に必要な項目を取捨選択し、目的別目次を用意しておく。新規作成者は適切な目次を選択し、選択された目次は作業領域に展開される。

2.4 プレゼンテーション編集インタフェース

(1) プレゼンテーション編集の流れ

本システムでは、再利用可能と判断されたスライドの URI を作業領域に展開された目次の各項目に紐付けながら、スライドを作業領域に格納する。

図 2 に開発したプレゼンテーション編集のためのユーザインタフェースを示す。①-A には、2.3 で選択された目次が展開される。また、目次の項目毎にスライド情報を格納でき、①-B で選択した項目に格納されたスライド一覧が閲覧出来る。②は検索機能を実現する部分、③は検索結果を表示する部分であり、検索結果のスライドの URI に紐付けられたサムネイル(③-A)や画像(③-B)を初めとする周辺情報(③-C,D,E)が表示される。また、④では検索以外のスライド格納機能やダウンロード機能を実現する。

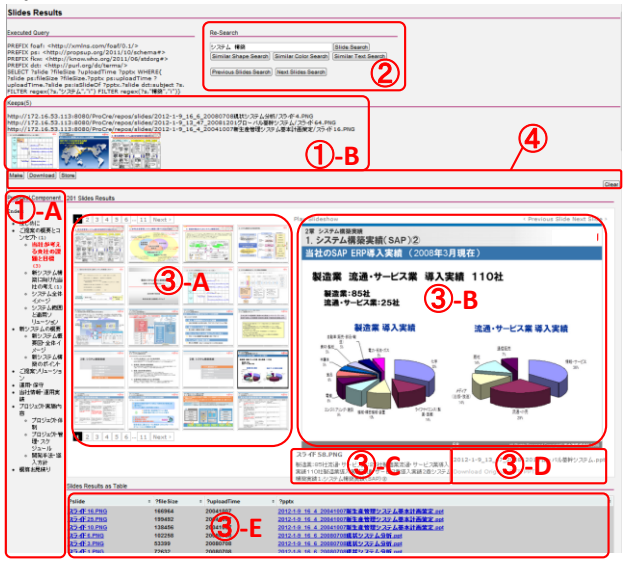

図2 ユーザインタフェース

(2) スライド情報の検索機能

本システムは、 入力した文字列をクエリとする全文検索機能、基本情報のメタデータを元に選択したスライドの前後のスライドを検索する機能、スライド類似性に関するメタデータを元に検索する機能の 3 つを備える。

2.5 ダウンロードインタフェース

編集作業により、作業領域に目次の項目と対応するスライドが紐付けられている。それらのスライドを目次の順序に合わせて一つのプレゼンテーションとしてマージし、プレゼンテーション作成者に提供する。

3. ケーススタディ

3.1 ケーススタディ概要

本ケーススタディでは、まず本システムを用いた既存プレゼンテーションスライド検索の性能評価について述べる。次に、実際に本システムを用いて、代表的なプレゼンテーションである提案書の作成を行い、現行の作成方法と比較することでシステムの有効性の評価を行った。

3.2 既存プレゼンテーション群

本ケーススタディでは、計 1632 枚のスライドを含む 17 つの既存プレゼンテーションファイルを再利用する対象とする。

3.3 個別のスライド検索性能の評価

まず、全文検索と類似度検索の検索性能評価を行う。検索対象スライドを「システム開発標準プロセス体系 SDEM に関するポンチ絵を含むスライド」とする。全プレゼンテーション内にこの条件を満たすスライドは図 3 に示す 6 スライドであった。

図3 検索対象スライド

全文検索のみで検索する場合と、全文検索と類似度検索を組み合わせた検索を行う場合の検索結果を表.1 に示す。全文検索においては、様々な検索クエリを送信するものの、スライド 5, 6 を検索するまでに、6 ステップを要する。一方、全文検索におけるクエリ id=1 の「SDEM」の検索結果のスライドをクエリとして類似度検索を用いた場合、テキスト類似度検索を行えば、基準とするスライドを id=1,2,4 とすれば、2 ステップ目で全 6 スライドを検索可能である。

表.1 全文検索による検索結果

Full Text Searcg		Search Result				
Query id	Query	num of results	correct slide id	Precision	Recall	F-measure
1	SDEM	9	1,2,3,4	0.44	0.67	0.53
2	標準プロセス体系	4	3,4	0.50	0.33	0.40
3	標準プロセス	10	3,4	0.20	0.33	0.25
4	標準　プロセス	29	1,2,3,4	0.14	0.67	0.23
5	開発標準	10	2,3	0.20	0.33	0.25
6	開発プロセス	8	1,2,4,5,6	0.63	0.83	0.71
7	開発　プロセス	56	1,2,3,4,5,6	0.11	1.00	0.19

表.2 類似度検索による検索結果

Similar Search			Search Result				
Query id	Query Slide id	Type of Similarity	num of results	correct slide id	Precision	Recall	F-measure
1	1	Shape	10	1,6	0.20	0.33	0.25
2	1	Color	10	1,	0.10	0.17	0.13
3	1	Text	23	1,2,3,4,5,6	0.26	1.00	0.41
4	2	Shape	10	2,4,5	0.30	0.50	0.38
5	2	Color	10	2,	0.10	0.17	0.13
6	2	Text	35	1,2,3,4,5,6	0.17	1.00	0.29
7	3	Shape	10	3,4	0.20	0.33	0.25
8	3	Color	10	3,	0.10	0.17	0.13
9	3	Text	10	1,2,3,4	0.40	0.67	0.50
10	4	Shape	101	2,3,4,5,6	0.05	0.83	0.09
11	4	Color	10	4,	0.10	0.17	0.13
12	4	Text	34	1,2,3,4,5,6	0.18	1.00	0.30

3.4 プレゼンテーション作成シナリオ

プレゼンテーションの作成シナリオを以下とし、基本的な目次を設定して作成する。

【被験者】IT ベンダー A 社　業務経験 16 年の社員
【提案形式】紹介資料(50 枚程度のプレゼンテーション)
【提案内容】個別受注生産の生産管理システムの再構築
【顧客要望】①納期短縮　②コストダウン

3.5 評価方法

被験者に 2 通りの方法によって新規プレゼンテーションの草案を作成してもらう。作成の流れをビデオで記録し、作成されたプレゼンテーションの草案を比較した。ただし、以下の 2 通りのプレゼンテーション作成はできるだけ事前知識の公平性を保つため、一週間の間隔を空け、作業の制限時間を 2 時間とした。

(作成方法-①) 現行の方法によるプレゼンテーション作成

17 つのプレゼンテーションを PC のデスクトップに置いた状態で、Microsoft PowerPoint2010 のみを用いて作成する。

(作成方法-②) 本システムを用いたプレゼンテーション作成

本システムに 17 つのプレゼンテーションを事前に蓄積しておき、本システムを用いて検索、抽出、作成を行う。

3.6 作成プロセスの比較

表.3 に 2 通りの作成方法に関する各過程の所要時間を示す。作成方法①の実作業時間は 109 分、作成方法②の場合は 87.5 分(自動統合処理の 8.5 分は除いた)であった。本システムを用いた作業時間は現行の方法に比べ、19.7%短縮された。

これの主な要因は 2 点考えられる。1 点目は、現行の方法では、ファイルを開閉する無駄と、既存プレゼンテーションファイル内のスライド順通りに再利用可能性を判断する必要が生じるが、本システムを用いる場合は、スライド単位で一元管理がなされている上、2.4.2 の各種検索の検索結果についてのみ再利用可能性を判断すればよく、処理スライド数は激減する。

2 点目は現行の方法では、抽出と組換えの過程が連続的であり、まず有益と判断したスライドを抽出し、その後それらのスライドを並び替えながら統合する。一方で、本システムは、検索の段階で、目次に対して紐づけを行うことで、抽出と組換え作業の並列化が行われている。その後の不要スライドの削除と修正プロセスにおいても時間が短縮できているのがわかる。

表.3 各プロセスの所要時間内訳 (分)

Process	Existing method	This system
Preprocess	5(Open all files)	0.5 (Select TOC)
Search	55	66
Extract		
Reassembly	21	
		8.5 (Automatic)
Delete and Modify	28	21
Total	109	96

3.7 作成されたプレゼンテーションの比較

先述した作成方法と作成段階(A.削除修正前、B.削除修正後)の観点から表.4 の 4 つのプレゼンテーションを評価する。

表.4 各提案書に含まれるスライド数(枚)

	Existing method①	This system②
A. Before modification	74	70
B. After modification	44	41

まず、各プレゼンテーションにおける提案目次の項目別のスライド数を図 4 に示す。修正後のプレゼンテーション①-B、②-B の分布が類似しているのに対し、修正前の収集を終えた段階の①-A は約半数の 35 枚が「第 4 章」に関するスライドである。一方で、「第 5 章」に関するスライドが収集されていなかった。比べて、②-A は比較的バランスよく収集されていることがわかる。

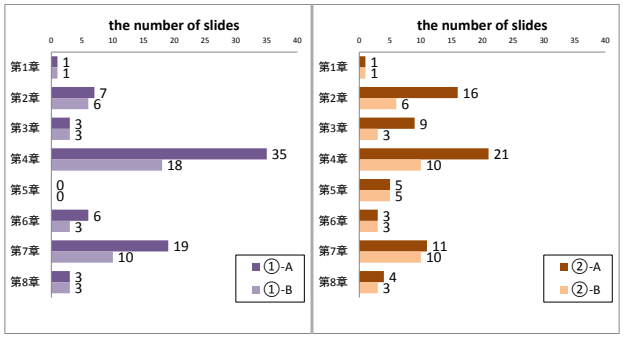

図1 目次の各項目に含まれるスライド数の違い(左図: ①-A, ①-B, 右図 : ②-A, ②-B)

次に、抽出元のプレゼンテーションファイル別に集計した表を図 5 に示す。現行の方法で作成された①-A, ①-B に関しては、7 ファイルから抽出されたスライドのみで作成される。一方で、②-A, ②-B に関しては、それぞれ 15 ファイル、12 ファイルから抽出されている。

最後に、再利用されたスライドの抽出元プレゼンテーションファイルにおける位置の一部(スライド番号 1～99) を図 6 に示す。横軸はプレゼンテーションファイルにおけるスライド番号を表す。①-A は数種類のプレゼンテーションから、連続的に抽出されているのがわかる。一方、②-A については、同一のプレゼンテーションファイル内においても、その抽出位置が前後にばらついていることがわかる。

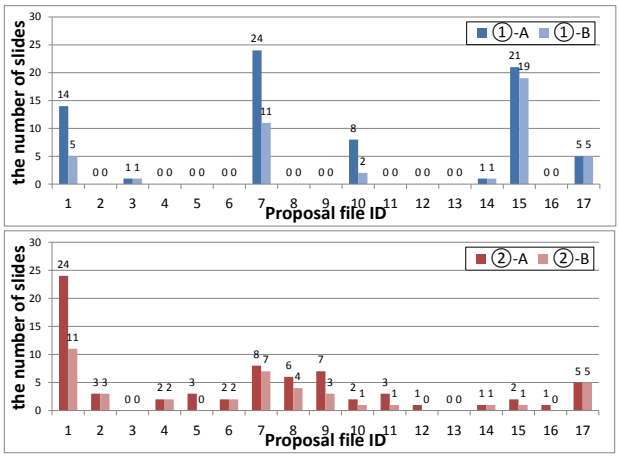

図2 各プレゼンテーションファイルから抽出されたスライド数
(上図 : ①-A, ①-B, 下図:②-A、②-B)

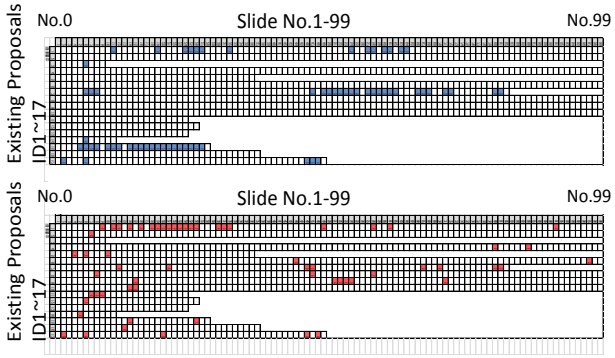

図3 抽出されたスライドの既存プレゼンテーションファイルにおける位置 (上図:①-A, 下図:②-A)

3.8 利用履歴に基づく各検索機能の比較

スライドの格納の直前の検索はそのスライドの発見に寄与したことを意味する。本システムはスライドの検索、格納等のアクションのログを記録している。ログの解析から、2.4.2 の各種検索回数およびスライドの格納に寄与した検索の回数を図 7 に示す。

検索プロセス全体の 7 割弱は全文検索であるが、スライドの格納に関しては 39%が全文検索によるものであり、35%がテキスト類似検索によるものであった。検索一回あたりの格納スライド数はテキスト類似検索が最も多く 2.0(枚)を上回ることがわかる。

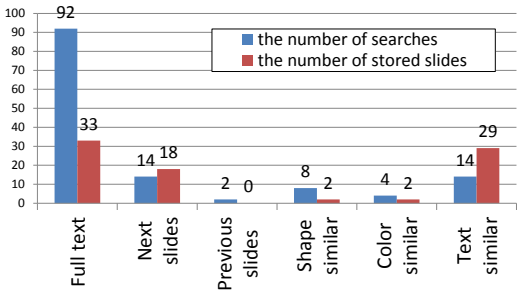

図4 6 種類の検索方法の比較

4. 考察

4.1 検索対象スライドへの前提知識の影響

検索対象スライドに対する知識が十分にある作成者は、関連する文字列が思いつくため全文検索のみで目的のスライドに到達できる可能性が高い。知識が乏しい場合は代替文字列が思い浮かばず、検索できない可能性がある。しかし、そのような知識が乏しい場合に対しても、3.3 の結果から類似度検索を用いれば有効な検索ができると考えられる。このようなクエリの支援の手段としては、オントロジーの利用も考えられるが、本研究では自動的に付与できるメタデータのみを対象としており、人手を用いたオントロジーの作成なしでもクエリの補完支援ができたことに意味がある。

4.2 各項目に対するスライド漏れの抑止効果

本システムは作業領域に展開されたプレゼンテーション目次に格納する形で、再利用するスライドを収集する。そのため作成者に対し、目次の項目に対する強い意識付けが行われるため、現行の方法による作成で生じた項目の欠落というミスは軽減できると考えられる。

4.3 元のプレゼンテーションファイルの多様性

現行の方法では、各プレゼンテーションに対してタイトルおよび数枚のスライドのみでシナリオとの合致性を判断し、プレゼンテーション全体の取捨選択を行うため、プレゼンテーションの後半は確認されないこともある。一方、プレゼンテーション②-A については、①-A では採用されなかった 9 つのプレゼンテーションファイル(id=2,4,5,6,8,9,11,12,16)からもスライドを抽出している。これらの提案内容はシナリオで指定された「個別受注生産」とは異なるが、今回提案するシステムと同様のパッケージ製品の説明に関するスライドや、詳細なプロジェクト体制図が含まれていた。また、図 6 から、②-A もついては同一ファイル内における隔たりを超えて様々なスライドが抽出されている。このように提案内容と直接的に依存しない部分で提案内容の異なる様々なプレゼンテーションファイルに含まれる有益なスライドが抽出できていることがわかる。

5. 結論

本研究ではプレゼンテーション作成支援システムを開発した。プレゼンテーションファイルを URI を用いてスライド情報単位で管理し、RDF を用いて適切な属性情報の付与や類似スライドの関連付けを行うことでスライド再利用効率の向上、とりわけ候補スライドの網羅的な検索を実現した。

また、メタデータとして情報を紐づける際に、自然言語処理、画像処理技術を用いることで、属人性を排した網羅的な関連づけ、属性情報の付与を実現した。

ケーススタディによって本システムを評価した結果、現行の方法に比べ、プレゼンテーション作成時間が 2 割程度短縮した。また、本システムを用いて作成されたプレゼンテーションは、現行の方法に比べ、広い検索領域から、適切なスライドを抽出して作成されていることがわかった。本システムにより記録されたログデータを分析することで、メタデータを用いた類似スライドの関連づけの有効性を示した。

以上より、効率的に多様なプレゼンテーションを作成できるという点で本プレゼンテーション作成支援システムは有効である。

参考文献

[神崎 2005] 神崎正英 : セマンティック・ウェブのための RDF/OWL 入門, 森北出版株式会社, (2005)

[Sivic 2003] J. Sivic and A. Zisserman : Video Google: A text retrieval approach to object matching in videos, in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on Computer Vision, pp. 1470–1477, (2003)

[Swain 1991] M. J. Swain and D. H. Ballard : Color indexing, International journal of computer vision, vol. 7, no. 1, pp. 11–32, (1991)

[Eishbein 2008] Jonathan M. Fishbein , Chris Eliasmith : Integrating structure and meaning: a new method for encoding structure for text classification, Proceedings of the IR research, 30th European conference on Advances in information retrieval, (2008)