# 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

47106692　笈田　佳彰
指導教員　稗方　和夫　准教授

Document information management system for proposal creation was developed. Proposals 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 proposal creation is reduced by around 20% using the system and the created proposal includes the more various slides which are included in different proposal files.

Key words: Metadata, RDF, Information retrieval, Document creating support

## 1　緒言

提案書とは提案活動を行う上で必須な文書情報であり、某 IT ベンダーの基幹システム提案チームは年間 400 件程度の提案書を作成する。提案書作成業務の中で、既存提案書の再利用は欠かせないプロセスであるが、100 枚～300 枚程度のスライドを含む PowerPoint ファイルであるため、目的スライドを検索し、再利用する際に、提案書ファイルの開閉や、提案書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。多様な顧客の多様な課題・要望に応じて柔軟に対応するには困難が伴う。

そこで本研究では、提案書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提案書の再利用効率の向上を目指した文書情報管理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、多様な提案書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　プロジェクト提案のための文書情報管理システム

### 2.1　文書情報管理システム概要

本研究で開発したシステム図を Fig. 1 に示す。開発したシステムは、web 上に構築され、大きく分けてレポジトリ部とインターフェイス部からなる。これより各インターフェイスを中心に説明する。

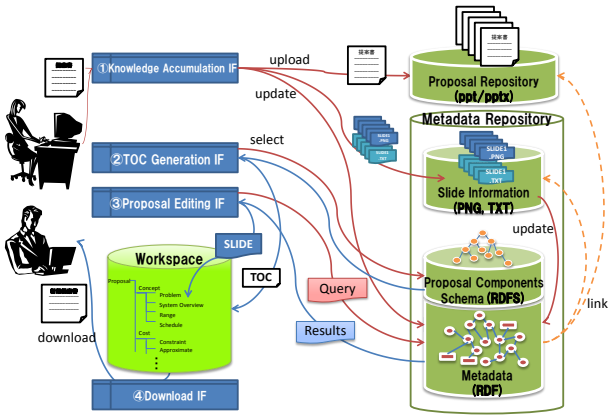


Fig. 1 System overview

### 2.2　知識蓄積インターフェイス

提案書を再利用しやすい形式に変換し、本文書情報管理システムに蓄積するためのインターフェイスである。

#### 2.2.1　文書情報蓄積粒度

本システムではスライド単位と提案書単位の 2 つの粒度で提案書文書情報を管理する。蓄積粒度の管理には、提案書をアップロードする段階で、提案書と提案書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。

#### 2.2.2　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、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。以下付与するメタデータの種類別に説明を行う。2.2.3 の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、2.2.4 の類似スライドに関するメタデータおよび提案書の標準項目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。

#### 2.2.3　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、提案書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。

#### 2.2.4　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、色のような情報に加え、テキスト情報も用いる。

**(1)形状情報に関する特徴量(Bag of visual words[1])**

スライド画像をスケール、回転に不変で、明暗にロバストな局所特徴量である SURF 特徴量の集合とみなし特徴ベクトル化を図る。まず、画像集合から、すべての SURF 特徴量を抽出し、それらを K-means クラスタリングすることで(本システムでは K= 500 に設定)、K 個のクラスタを代表するベクトル(=Visual Words)を得る。次に、各画像から同様に SURF 特徴量を抽出し、最近傍に位置する Visual Words に投票することで、Visual Words を横軸とした正規化されたヒストグラムを得る。

**(2)色情報に関する特徴量(Color Histogram[2])**

RGB256 階調の場合、16,777,216 色表現可能である。類似の幅を持たせるため、4 階調に減色し、64 次元のヒストグラムを特徴量とした。

**(3)テキスト情報に関する特徴量(Bag of words[3])**

スライド内に含まれる文字列をスライドに含まれるオートシェイプ毎に抽出し、かつ体裁のために含まれる不要なスペースを除去した後、形態素解析を行い、提案書群内に含まれる全ての単語を抽出する。その中でストップワードを除いて出現頻度の高い上位 1000 語を横軸とし、各スライドに対して、その 1000 語に関する出現頻度を計算し、正規化されたヒストグラムを得る。

**(4 )バタチャリヤ係数を用いた類似度計算**

二つのヒストグラム ***I, M*** に対して、式(1)を用いて類似

度の計算を行う。3 つの特徴量についてそれぞれ閾値を定め、閾値を超えるスライド間に、どの特徴量の観点から類似するかを意味するメタデータを付与し関連付けた。

$$\mathbf{BC}(\boldsymbol{I},\boldsymbol{M}) = \sum_{\mathbf{j=1}}^{\mathbf{n}} \sqrt{I_j M_j} \qquad (1)$$

### 2.3 目次生成インターフェイス

本インターフェイスは提案書作成において、新規提案書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提案書標準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 提案書編集インターフェイス

#### 2.4.1 本システムを用いた提案書編集の流れ

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 2.3 で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。

Fig. 2 に開発した提案書編集のためのユーザーインターフェイスを示す。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。

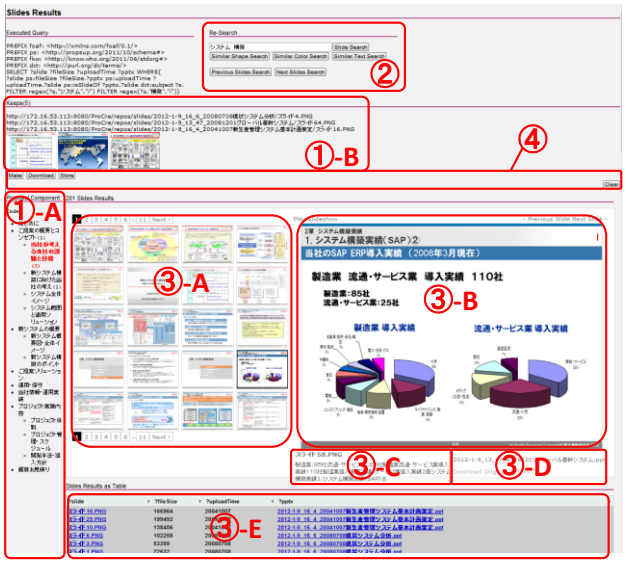


Fig. 2 User Interface

#### 2.4.2 スライド情報の検索機能

本システムは以下の検索機能を有する。

- 入力した文字列をクエリとする全文検索機能
- 2.2.3 で付与した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
- 2.2.4 で付与したメタデータを元に選択したスライドの類似スライドを検索する機能

#### 2.4.3 ダウンロードインターフェイス

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提案書としてマージし、提案書作成者に提供する。

## 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本文書情報管理システムを用いて既存提案書を管理し、スライド検索の性能評価を行う。次に実際に本システムを用いて提案書作成を行い、現行の作成方法と比較することでシステムの有効性を評価する。

#### 3.1.1 既存提案書群

本ケーススタディでは、計 1632 枚のスライドを含む 17 つの既存提案書を再利用する対象とする。

#### 3.1.2 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。目的スライドを「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。全提案書内にこの条件を満たすスライドは Fig. 3 に示す 6 スライドであった。


Fig. 3 Target slides

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 Table.1 に示す。全文検索においては、様々な検索クエリを投げるものの、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。

Table.1 Search results by full text search

| 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 |

Table.2 Search results by search based on similarity

| 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 |

### 3.1.3 提案書作成シナリオ

提案書作成シナリオを以下のように設定する。
【作成者】IT ベンダー A 社　ベテラン社員
【提案先(顧客)】一部上場　重工業 Q 社
【提案形式】ご紹介資料
【提案内容】個別受注のため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
作成する提案書の目次を Table.3 に示す。

Table.3 Contents for proposal

| Headings | | Subheadings | |
|---|---|---|---|
| 1 | はじめに | | |
| 2 | ご提案の概要とコンセプト | 1 | 当社が考える貴社の課題と目標 |
| | | 2 | 新システム構築に向けた当社の考え |
| | | 3 | システム全体イメージ |
| | | 4 | システム範囲と適用ソリューション |
| 3 | 新システムの概要 | 1 | 新システム概要図 |
| | | 2 | 新システム構築のポイント |
| 4 | ご提案ソリューション | | |
| 5 | 運用・保守 | | |
| 6 | 当社情報・実績 | | |
| 7 | プロジェクト実施内容 | 1 | プロジェクト体制 |
| | | 2 | プロジェクト管理 |
| | | 3 | 開発手法・導入方針 |
| | | 4 | システム構築スケジュール |
| 8 | 概算お見積り | | |

### 3.1.4 評価方法

実際に某 IT ベンダーの提案業務経験 16 年のベテラン社員に以下の 2 通りの方法によって新規提案書の草案を作成してもらう。作成の流れをビデオで記録し、作成された提案書の草案を比較した。ただし、以下の 2 通りの提案書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た。また作業の制限時間を 2 時間とした。

(作成方法-①) 現行の方法による提案書作成

17 つの提案書を 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提案書作成

本文書情報管理システムに 17 つの提案書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。

### 3.1.5 作成プロセスの比較

Table.4 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、既存提案書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の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。

Table.4 Time required for each process (min)

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

### 3.2 作成された提案書の比較

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 Table.5 に示す 4 つの提案書を評価する。

Table.5 The number of slide in each proposal

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

まず、各提案書における提案目次の項目別のスライド数を Fig. 4 に示す。修正後の提案書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「ご提案ソリューション」に関するスライドである。一方で、「運用・保守」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。

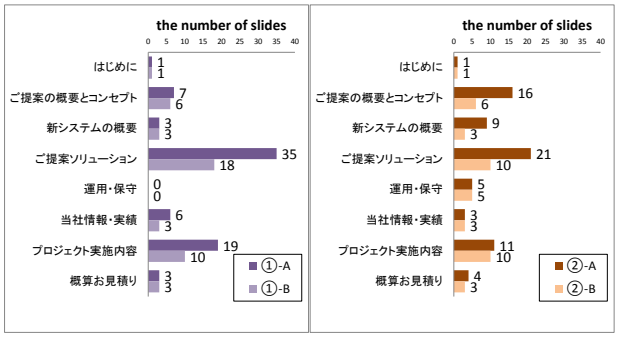


Fig. 4 The number of slides included in each item (left：①-A, ①-B, right：②-A, ②-B)

次に、抽出元の提案書ファイル別に集計した表を Fig. 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。

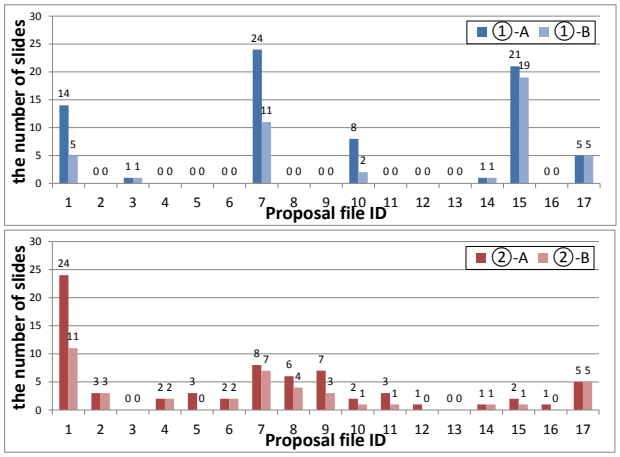


Fig. 5 The number of slides extracted from each proposal (above：①-A, ①-B,　below:②-A、②-B)

最後に、再利用されたスライドの抽出元提案書における位置の一部(スライド番号 1～99) を Fig. 6 に示す。横軸は提案書におけるスライド番号を表す。①-A は数種類の提案書から、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、同一の提案書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。

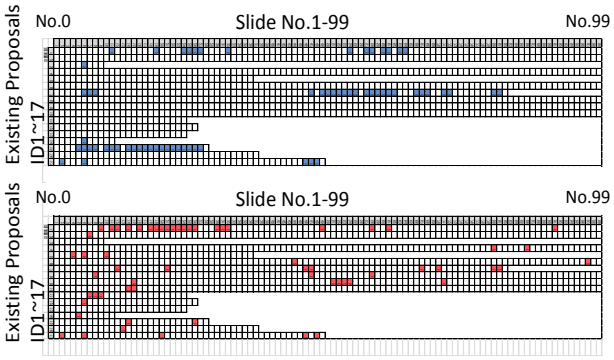


Fig. 6 The position of the extracted slides in existing proposals (above:①-A, below:②-A)

### 3.3 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。本文書情報管理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。記録されたログを解析し、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 Fig. 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、格納に寄与したスライド数はテキスト類似スライドと 4 枚しか変わらない。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 を上回ることがわかる。

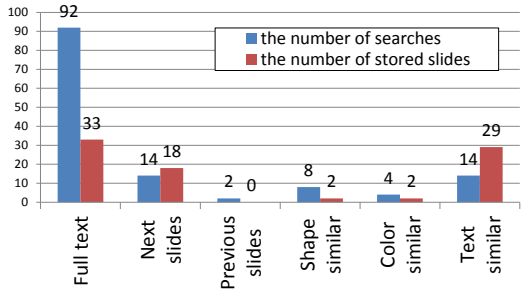


Fig. 7 Comparison of 6 types of searches

## 4 考察

### 4.1 目的スライドに関する知識の有無と本システムの有効性

目的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、検索が頓挫する可能性がある。しかし、そのような知識が乏しい場合に対しても、3.1.2 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について

本システムは作業領域に展開された提案書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 スライド抽出元の提案書の多様性について

現行の方法では、各提案書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、提案書全体の取捨選択を行うため、提案書の後半は確認されないこともある。一方、提案書②-A については、①-A では採用されなかった 9 つの提案書(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、Fig. 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提案書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 結言

本研究ではプロジェクト提案のための文書情報管理システムを開発した。提案書を 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文書情報管理システムの評価した結果、現行の方法に比べ、提案書作成時間が 2 割程度短縮した。また、本システムを用いて作成された提案書は、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、効率的に多様な提案書を作成できるという点で本文書情報管理システムは有効である。

### 文献

1) 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, 2003, pp. 1470–1477
2) M. J. Swain and D. H. Ballard：Color indexing, International journal of computer vision, vol. 7, no. 1, 1991, pp. 11–32
3) 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, March 30-April 03, 2008