O-012

図形集合に基づくプレゼンテーション文書検索
Presentation Document Search Based on Clustered Figures

小柳 隆人†1 角田 啓介†1 鈴木 健太郎†1 宮下 直也†1 箕浦 大祐†1
Ryuto Koyanagi　Keisuke Tsunoda　Kentaro Suzuki　Naoya Miyashita　Daisuke Minoura

1. はじめに

本稿ではプレゼンテーション資料において，類似の図形集合を含む文書を検索する手法を提案する．

企業活動において，プレゼンテーション資料は，内部向け，外部向けに関わらず，広く説明および検討に用いられている．プレゼンテーション資料においては，文章だけでなく，図表等を用いて，視覚的に閲覧者・聴講者が理解しやすいよう工夫される．企業におけるプレゼンテーション資料は，いわゆる発表に使用されるだけでなく，説明のための資料として使われるケースが多い．そのためその工夫された図表は抜粋され，一部を修正して流用される．例えば閲覧者に応じて，内部向け，顧客向け等，閲覧者に合わせて既存の資料を変形，修正し使い回すケースなどがある．また，閲覧者側も抜粋や修正された資料のみを受領することがある．

これらの状況において，自分が作成した資料が他の目的でどのように使用されたか知りたい，あるいは受け取った元となる資料を探したいというニーズが存在する．しかし，一般的なプレゼンテーションアプリケーション，例えば，Microsoft PowerPoint[1], Google スライド[2], LibreOffice[3]等において，そのような図形を検索する機能は，本稿記載時点において存在しない．富士通ではプレゼンテーションから抽出した，言語情報を用いて，スライドの情報を抜き出し，特徴語をもとにスライドを検索する技術が検討されている[4] が，公開されていない．また大阪大学は，執筆中の UML 図を元に類似の UML 図をレコメンドする開発支援ツールを提案している[5]が，UML は開発という特定の用途に特化したものであり，使用される状況は限定される．

そこで，本研究はプレゼンテーション資料におけるその工夫された図形を元に類似の図形を含むプレゼンテーションを探すことを目的とする．この目的を満たせれば，資料がどのように流用され使用されているかを把握したり，抜粋された資料の元となった資料を探し，元となった資料が作成された経緯を確認することが可能となる．

2. 関連技術

2.1 全文検索技術

文書を検索する技術として，テキスト情報を検索する全文検索技術が存在する[6]．全文検索技術では，一般に入力したテキストに対し，マッチするテキストを含む文書を検索することができる．全文検索技術では，辞書が適切にメンテナンスされており，探したい類似の図形に紐づくテキストが文書に存在する場合は，類似の図形を含むプレゼンテーション文書を発見することができる．一方，その類似の図形に紐づく文章がないケースや変更されているケース，特徴的な語を含んでおらず検索ノイズが大量にでるケース等では目的の文書を発見することが困難である．

†1 NTT コムウェア株式会社
NTT COMWARE CORPORATION

2.2 画像検索技術

文書を検索するにあたり，スライド，あるいは図形の集合を画像に変換して検索する方法が考えられる．ここでは画像を検索する方法としてハッシングと特徴点マッチング，機械学習による画像分類を例として挙げる．

2.2.1 ハッシング

ハッシングでは，画像を特定のサイズに圧縮し，その輝度等の情報を用いて，画像を画像同士の距離を計算可能なハッシュ値に変換する[7]．そして検索クエリとして，画像が与えられた時，同様に変換されたクエリ画像のハッシュ値と距離が近いハッシュ値を持つ画像を検索結果とする．この手法は，検索速度が早く類似画像検索にてよく利用される手法である．しかし，本稿の目的においては，資料が使い回されることを想定しており，この場合，その資料の余白に合わせて図形集合が変形されるケースが多い．この変形が発生すると画像のハッシュ値が大きく異なってしまい，目的とする文書を発見することが困難と考えられる．

2.2.2 特徴点マッチング

画像に含まれる物体を検索する際に特徴点をマッチングし，同一の物体が存在すると判定する手法がある[8]．特徴点マッチングでは画素ごとに位置をずらし，その要素の差分が大きい箇所を特徴点として，その特徴がマッチする割合が大きいほど近い画像として検索可能である．しかし，本稿の検索対象であるプレゼンテーション文書は写真等のラスタ画像ではなく，図形を組み合わせて作られたベクタを元にプレゼンテーションソフトウェアが描画した画像であり，後述の評価で使用したデータにおいては，プレゼンテーション文書内に含まれる図形のうち，6 割以上が「長方形」であった．そのため同型の図形が多く，似た特徴が多く算出され，目的の文書を発見することが困難と考えられる．

2.2.3 機械学習による画像分類

機械学習を用いて画像を特定のクラスに分類する[9]ことで画像の検索に利用できる可能性が考えられる．本手法を用いると，例えば，猫の画像と猫がいない画像を学習させて，入力されてきた画像に猫を含むか否かを判定することが可能である．しかし，今回の目的は検索であるため，同一の素材を含む画像の分だけのクラスを分類をする必要がある．このとき，教師として使えるデータはごく少量にも関わらず大量のクラスに分類する必要があり，今回の目的では適切でないと考えられる．

3. 提案手法

本稿では以下の仮説および定義を元に図形集合を検索する手法を提案する．

仮説 1：プレゼンテーション文書のスライドに存在する図形は一部の例外（テキストによる修飾等）を除き，単独では意味を持たない．このとき，図形の集合により意味をなす単位が存在し，以降「素材」と呼ぶこととする．

仮説 2：素材が流用され，プレゼンテーション文書が作成されるとき，素材は拡大，縮小，図形の欠落，図形の追加を伴う．これに頑健な特徴を抽出することで類似の図形集合を含む文書を発見することができる．

提案手法は大きく以下の 3 STEP（図 1）からなる．

3.1)素材の切り出し

3.2)素材からの特徴抽出

3.3)クエリの生成と検索

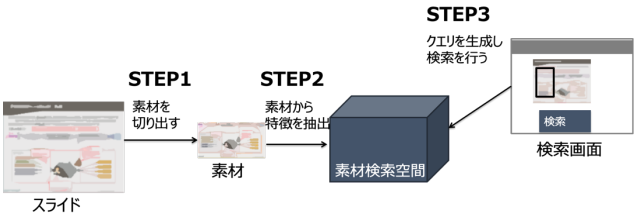

図 1 提案手法概要

なお，今回は，プレゼンテーション文書として，ECMA-376[10]規格に沿う Microsoft 社の Microsoft Powerpoint ソフトウェアにて作成された.pptx ファイルを対象とし，実装・評価を実施した．

3.1 素材の切り出し

素材の切り出しを行うための前準備として，共有フォルダをクロールし.pptx ファイルを取得する．その.pptx ファイルを zip 解凍，各スライドの xml ファイルから図形が持つ情報を取得しておく．

素材を用いて類似の図形集合を含む文書を発見するためには，スライドに存在する図形を集め，素材として切り出す必要がある．本 STEP では，各図形間の距離を元にクラスタリングを行い，距離が近いものを同一の素材として切り出しを行う．これを図 2 に示す．今回は，クラスタリング手法として，凝集度階層クラスタリング(Agglomerative Hierarchical Clustering)を用い，クラスタ間距離の算出方法は ward 法を使用した[11].

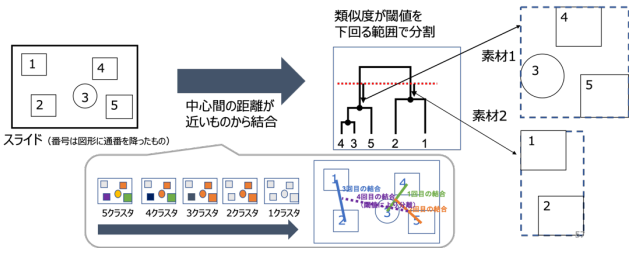

図 2 素材切り出し手法

3.2 素材からの特徴抽出

本 STEP では，前 STEP で切り出した素材に対し，図形の形とその素材における中心からの位置関係を用いて，図形ごとに下記 4 つの特徴を算出する．これを図 3 に示す．

特徴 1．Geometry（図形の形を表す番号）：ここでは，ECMA-376 にて定義されている preset_geometry というパラメータを元に番号を割り振る．例えば「長方形」の図形は，preset_geometry が「rect」であるため，「rect」に割り当てた番号「144」を付与する．ここで見た目上同型となる形（例えば円柱が描画される「can」と磁気ディスクが描画される「magnetic」等）については，正規化を行い，同じ番号を付与する．

特徴 2．Distance（中央からの距離）：素材の中心から最も遠い点までの距離を 1 として，その図形の中央座標までの距離を算出する

特徴 3．Theta（角度）：素材の中心を原点としたときの図形の中心までの角度を算出する

特徴 4．Size（サイズ）：素材全体の面積を 1 としたときのその図形が占有する面積を算出する．本来であれば，図形の実際の面積は，その図形の種類により異なり，例えば，「上カーブ矢印」など求めることが困難なものがある．そのため，ここでは図形が持つ情報である，x 座標の位置，y 座標の位置，幅，高さ，の情報を元に図形を長方形とみなして近似面積（サイズ）を算出する．この近似面積は実際の面積とはかけ離れている値になることがあるが，後述する検索方式によって，同一でない図形については，検索時に一致判定を行わないため，影響は微小となる．

この手法では素材内での位置関係を用いることで，拡大・縮小による特徴量の変化を低減する．また原点を 4 隅ではなく，中央としたのは，theta を 360 度とすることで後述する検索方式によって閾値を調整しやすくするためである．これを全ての図形，全てのファイルに対して実施し，その結果を 1 つの検索リソースファイル（検索に利用する数値のみの行列ファイル）に出力する．

3.3 クエリの生成と検索

3.2 までで検索リソースとなるデータを生成する前処理部分を実施した．ここでは，ユーザが実際に検索を行う部分の処理を行う．検索の流れは以下の通りとなる．ユーザは検索システムに検索したい図形集合が含まれる画像ファイルをアップロードする．検索システムはアップロードされたファイルの展開とスライドの画像化を行い，スライド画像を画面に表示する．ユーザは検索したい図形を含むページを選択し，マウスのドラッグ操作にて図形集合の範囲を指定し，検索を押す．このとき必要に応じて検索の閾値を調整する．検索システムは類似の図形集合を含むファイルの情報を検索結果として，表示する．検索画面のイメージを図 4 に示す．

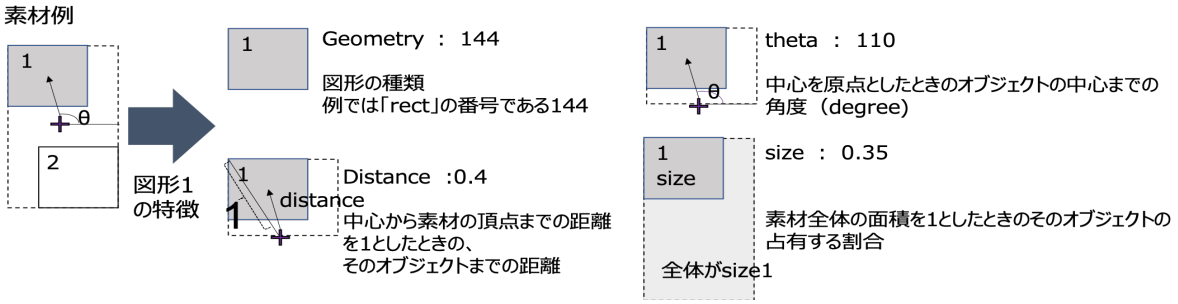

図 3 特徴抽出手法

検索の手法は以下の通りである．検索システムはユーザが選択したページおよび座標情報と閾値の情報を受信し，3.1 と同じ手法で選択された範囲に存在する図形が持つ情報を取得する．検索システムは選択された範囲を 1 素材と見なし，3.2 と同じ手法で各図形の特徴を算出する．システムは，受信した閾値の情報を元に同一とみなすレンジを算出する．例えば，オブジェクト 1 の theta が 70，閾値が 30 であった場合，theta のレンジは，40～100 となる．Distance，Size についても同様にレンジを算出する．これを範囲内の全図形に対して生成し，それぞれの図形に対し，Geometry が一致し，Distance，Theta，Size が閾値内に収まる図形を検索リソースファイルから探索し，一致するものにフラグを立てる．これを繰り返し，最終的にフラグが立った図形を最も多く含む素材上位 n 件を検索結果として，その素材を含むプレゼンテーション，スライドの情報を表示する．これを図 5 に示す．

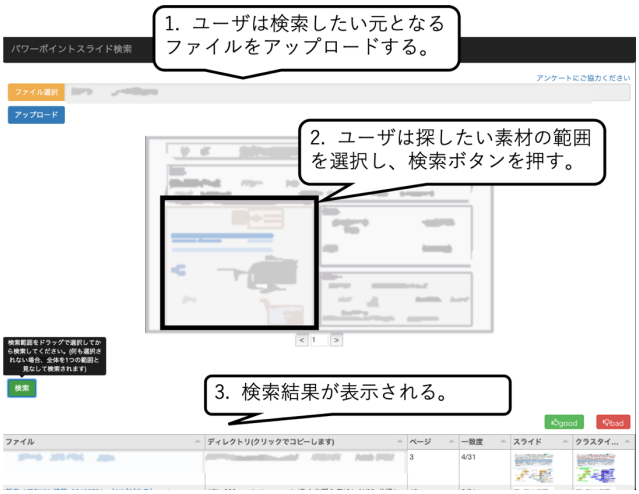

図 4 検索画面イメージ

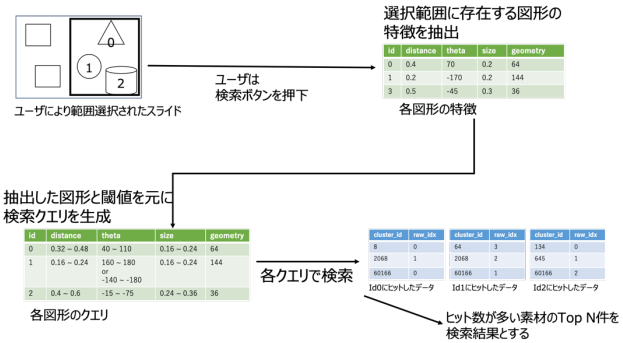

図 5 検索手法

4. 評価と考察

4.1.1 評価方法

提案手法の有効性を判断するため，検索用スライドを提示し，目的に設定したスライド（類似または完全に一致するスライドを持つファイル）を探す問題を計 5 問用意し，各問題 5 分間の制限時間内に発見できるかを 5 名のユーザにてテストを行い，テスト実施後に使いやすさに関するヒアリングを実施した．テストにて利用した検索リソースのデータ量は，.pptx ファイル数：6086，スライド数：21449，素材数：39232，オブジェクト数：352760 である．問題の概要は以下の通りである．

問題 1：全く同じシステム構成図を探す

問題 2：一部のオブジェクトが異なるスケジュール表を探す

問題 3：同じパンフレットを探す

問題 4：Sprint5 の計画資料から Sprint1 の計画資料を探す(類似資料の存在が推定される資料)

問題 5：進捗報告資料から他グループが作成した進捗報告資料を探す(類似資料の存在が推定される資料)

検索時には，閾値として，同一とみなせる距離の範囲（デフォルト：±20%），同一と見なせる面積の範囲（デフォルト：±20%），同一とみなせる角度の範囲（デフォルト：±30°）をユーザが指定できるようにした．

4.1.2 評価結果

問題 1 から問題 5 において，問題 1～4 は 5 名中 4 名が正解，問題 5 は全員が正解したため，正答率は 84%であった．それぞれの問題の正解発見までの検索回数の平均および最大・最小値を図 6 に示す．正解の発見までの検索回数のうち平均検索回数は 3.55 であった．

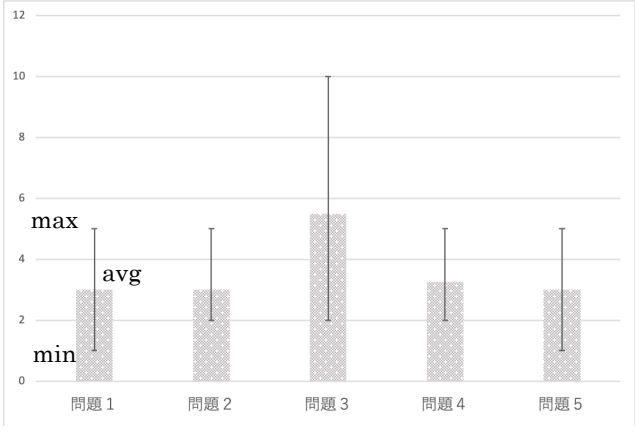

図 6 正解発見までの検索回数

4.1.3 考察

評価では，80%以上のユーザが目標とするファイルを発見することができた．また正答者の目的ファイル発見までの検索回数が平均 3.55 回であり，企業内で共有フォルダ等をやみくもに探すよりは早く目的のファイルを発見できると考えられる．類似手法が発見できなかったため，既存手法との比較は出来ないが，類似の図形を含むプレゼンテーションの発見に一定の有効性があると考える．

3.1)素材の切り出しについて，使いやすさに関するヒアリングでは，本システムが切り出した素材がどのような形なのかがわからないという意見が複数あげられた．そのため切り出し手法については評価者の想定と合わなかったものが多く改善の余地がある．

3.2)素材からの特徴抽出，3.3)クエリの生成と検索について，評価時のユーザのクエリを確認したところ，クエリとシステムが分割した素材の範囲が完全一致していないが正解ができていたものも多かった．一致していなかったが，類似の素材を含むスライドを発見できたことから特徴抽出手法および検索手法においては一定の有効性があると考察される

ユーザ指定出来るようにした検索の閾値については，探したい文書が指定した閾値からどれくらいずれているかをユーザが判断できないため，有効に活用できたユーザがおらず，有効な手段であるとは言えなかった．

3.1)素材切り出し手法の問題としては，主に下記２点が考えられる．

１）全ての図形を結合した後に分割するため，縦・横に長い図形を結合してしまったときに，重心がずれてしまい，折れ曲がった素材が切り出されてしまうケースがある．

２）現在の手法では個々の図形から全体を組み上げていくように切り出しているが，人間がスライドを見たときには，全体を俯瞰したあとに，意味の区切りを見ていくといったように，結合順序が逆となっている．また，素材の認識単位については個人差が大きい．適当なスライドをあげ，素材の切り出しについて５名にアンケートをとったところ，切り出し方が一致したのは２名のみであり，一致した２名もスライド全体を１素材と回答したものであった．

本稿で提案した仮説については，下記知見が得られた．

仮説１:図形の集合により意味をなす単位が存在し，「素材」と呼ぶこととする．

⇨ 評価にて類似の図形集合を含む文書を 80%以上のユーザが発見できたことから，一定の妥当性があると考えられる．一方，単位については個人差が大きく，同じ集合においても素材が一意に定まらないと言える．

仮説２: 素材が流用され，プレゼンテーション文書が作成されるとき，素材は拡大，縮小，図形の欠落，図形の追加を伴う．これに頑健な特徴を抽出することで類似の図形集合を含む文書を発見することができる．

⇨ 評価にて類似の図形集合を含む文書を 80%以上のユーザが発見できたことから，本仮説は有効と考えられる．

5. おわりに

本稿では，プレゼンテーション資料における，類似の図形集合を含む文書を検索する手法を提案した．また，ユーザ評価を通じて類似文書および同一文書の発見において，提案手法の一定の有効性を確認した．

今後の展開としては，素材の切り出し手法の改善が挙げられる．さらに多数のユーザ評価を行い，その結果に近づけることが出来るような素材の切り出しのための特徴抽出手法の検討とパラメータ調整を行うことで，検索精度の向上が見込まれる．

参考文献

[1] Microsoft Corporation: Microsoft PowerPoint, <https://products.office.com/ja-jp/powerpoint>(参照 2019−04−22)

[2] Google LLC: google スライド, <https://www.google.com/intl/ja_jp/slides/

[3] The Document Foundation: LibreOffice The Doc about/>(参照 2019−04−22) ument Foundation,<https://ja.libreoffice.org/>(参照 2019−04−22)

[4] 馬庭伸栄，粟津正輝，岡田伊策ほか: AI 技術を活用した SE 変革の実践, *雑誌 FUJITSU*, Vol.68, No.6, pp.75-83(2017)

[5] 松下誠, 早瀬康裕, & 井上克郎: 状況に応じた設計情報の再利用を支援する UML 図の自動推薦ツール. *電子情報通信学会技術研究報告. SS, ソフトウェアサイエンス*, *109*(456), pp.37-42. (2010).

[6] The Apache Software Foundation: Apache Solr Ref Guide,<https://www-us.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-7.7.pdf>(参照 2019−04−22)

[7] Kozat, S. S., Venkatesan, R., & Mihçak, M. K..: Robust perceptual image hashing via matrix invariants, *In 2004 International Conference on Image Processing*, 2004.(ICIP'04.) ,Vol. 5, pp. 3443-3446, IEEE.(2004).

[8] Baumberg, A.: Reliable feature matching across widely separated views, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. (CVPR 2000), (Cat. No. PR00662), Vol. 1, pp. 774-781, IEEE.(2000).

[9] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P.: Supervised machine learning: A review of classification techniques, *Emerging artificial intelligence applications in computer engineering*, Vol.160, pp.3-24(2007).

[10] Ecma International: Standard ECMA-376,<https://www.ecma-international.org/publications/standards/Ecma-376.htm>(参照 2019−04−22)

[11] Ward, Jr, J. H.: Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, Vol.58, No.301, pp. 236-244 (1963).