社内文書検索システム（3）
-視覚印象検索のための領域レイアウト抽出-

岡城純孝　　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

近年，企業のオフィスなどでは，PC環境の普及により電子的な社内文書が大量に作成されている．プレゼンテーション作成ソフトなどを使って作成されたこれらの社内文書は，図や写真などの画像，グラフや表などを含むものが多い．人が文書を閲覧した場合，文書中の語彙(キーワード)だけでなく説明図に使われた図形の形状や色，それらのレイアウトなども記憶に残っている．社内文書検索においては，以前に自分自身が作成した文書や他人が作成して一度は見たことがある文書について見た目の大まかな記憶に基づいた検索を行いたいというニーズがある[1]．

本稿では，視覚印象の一つである文書の領域レイアウトの分析・抽出とその領域特徴を用いた視覚印象検索について提案する．このような検索システムを用いることにより，ユーザは文書に対するあいまいで断片的な記憶をもとにその文書を検索する再発見型検索が可能となる．

2 領域特徴抽出

社内文書のスライドに含まれる部分領域の種別(テキスト領域，図領域など)とその配置(位置，大きさ，形状)に関する特徴を領域特徴と呼ぶ．筆者らはまず，「社内文書は，自然言語で記述される“テキスト領域”と，図／表／グラフ／イメージなどで記述される“図表領域”に大別され，それらの配置が印象として残りやすい」という仮説を立てた．このような領域特徴を抽出するためには，見た目にひとかたまりとなるテキスト領域および図表領域にスライドを分割する必要がある(図1参照)．

スライドをテキスト領域および図表領域に分割するには，次のような課題が存在する．

課題1：テキスト領域だけでなく図表領域にも文字列が存在するため，単純に文字列部分をテキスト領域と判定できない

課題2：テキスト領域および図表領域の構成部品のレイアウトは複雑かつ多様である

課題1に対して筆者らは，「テキスト領域を構成するテキストボックスは，1つのスライド内での数が少なく，ボックス内の文字サイズが大きくて文字数が多い(つまり，面積が大きい)．一方，図表領域を構成するテキストボックスは，1つのスライド内での数が多く，ボックス内の文字サイズが小さくて文字数が少ない(つまり，面積が小さい)．」と仮定した．この仮定に基づき，筆者らは，スライドごとにテキストボックスの面積ヒストグラムを生成し，その分布からテキスト領域を構成するものと図表領域を構成するものとに分類する，というアプローチを取った．

課題2に対しては，「意味的に関連性の高い部品は互いに近くに配置され，それらによって作られたひとかたまりの部品群がそれぞれ識別できるように間を空けて矩形状に配置される」と仮定した．この仮定に基づき，部品の最小外接矩形(MBR)を生成し，それらの見た目の距離に基づいて領域を生成する，というアプローチを取った．

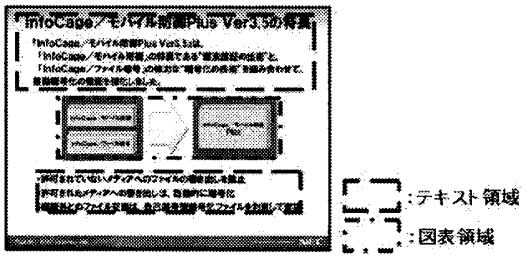

図1　テキスト領域と図表領域

以下に，上述のアプローチに則った部分領域抽出手順を述べる．

(1) 各スライドに含まれる部品を抽出する．
(2) 抽出した部品を，テキストを含む部品／含まない部品に分類し，テキストを含まない部品を図表領域の構成部品とする．
(3) テキストを含む部品について面積ヒストグラムを生成し，最頻値の面積より大きな面積を持つ部品をテキスト領域の構成部品，最頻値の面積以下の部品を図表領域の構成部品に分類する．
(4) テキスト領域および図表領域を構成する部品についてそれぞれ，部品のMBRが重なりを持つものを統合して新たなMBRを生成する．さらに，MBRの視覚印象距離(後述)がしきい値以内の部品を統合して新たなMBRを生成する．しきい値には，1つのスライドに含まれる任意の2つのMBRのすべての組み合わせの距離の平均値を用いる．

以上の手順により生成したMBRを，テキスト領

An In-house Documents Retrieval System (3) -Region Layout Extraction for Visual Impression-based Retrieval -
Sumitaka OKAJO, Katsushi MATSUDA
Service Platforms Research Laboratories, NEC Corporation

域および図表領域とする．

筆者らは，視覚印象距離を図 2 のように定義した．視覚印象距離によれば，2 つの部品の MBR の互いに向かい合う辺の距離が近いほど，さらに，それら 2 つの辺を辺に平行な軸に射影したときの重なりが大きいほど距離が小さくなる．

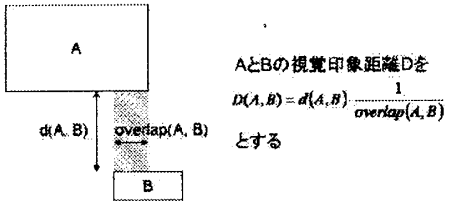

図 2　視覚印象距離

次に，抽出した各部分領域から特徴量を抽出する．本稿では，領域特徴量として領域種別(テキスト領域，あるいは図表領域)に加え，重心位置，面積，縦横比を抽出することとする．

3 ブロックパターンインデックス

類似文書を検索する際の処理時間を短縮するため，検索対象文書から抽出した各部分領域の位置，大きさ，形状の組み合わせのパターンを抽出元文書に対応付けたインデックスを作成し文書の絞り込みを行う．

まず，スライドを 3×3=9 個の等しいブロックに分割する．ブロックの分割に際しては，

- スライドには全体に満遍なくテキストや図表が配置される
- スライドの中心と，その上下左右という位置関係が印象に残りやすい

と仮定し，このような分割方法を採用した．次に，9 個のブロックから任意のブロックを選んだ場合に矩形となるブロックの組み合わせのパターン(計 36 パターン)と，領域種別をインデックスのキーとして部分領域を登録する．

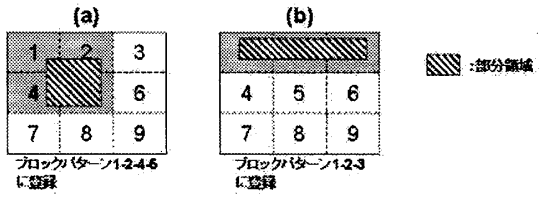

図 3　部分領域とブロックパターン

例えば，図 3(a)の部分領域はブロックパターン 1-2-4-5 に，図 3(b)の部分領域はブロックパターン 1-2-3 にそれぞれ属する．

本インデックスの 1 レコードは，ブロックパターン，領域種別，領域 ID から構成され，ブロックパターンと領域種別の組み合わせをキーとして領域 ID を得ることができる．また，各部分領域の詳細な特徴量である重心座標，面積，縦横比および，その部分領域を含むスライドについては，領域 ID をキーとする別のテーブルに格納する．このテーブルを参照することにより，領域 ID からそれらの値を得ることができる．

従来から空間インデックスに関しては R-Tree やその派生木が数多く研究されている[2]．しかし，これら従来方式はデータの登録や削除などインデックス管理が複雑であり，また，本研究では厳密な近傍検索や範囲検索は必要ないので，本インデックス方式で必要十分であると考えている．

4 類似文書検索

文書を検索する際には，ユーザは記憶にあるスライドの部分領域のレイアウトを検索クエリ(領域特徴クエリ)として入力する．続いて，2 節で述べた検索対象文書からの抽出と同様にして領域特徴クエリから領域特徴量を抽出する．次に，抽出した領域特徴量を用いて 3 節のブロックパターンインデックスを参照し，領域特徴クエリに含まれる部分領域ごとに，インデックスを検索し一致するスライドを取得する．次に，領域特徴クエリに含まれるすべての部分領域を持つスライドを特定する．具体的には，各部分領域ごとに取得したスライドの論理積を取ることで特定することができる．

最後に，絞り込んだスライドと領域特徴クエリとの詳細な類似度(スライド類似度)を計算する．スライド類似度の計算には，各々対応する部分領域の類似度である部分領域類似度の平均を用いる．部分領域類似度の計算には，部分領域の特徴量から得られる特徴ベクトルのなす角 θ によるコサイン尺度を用いることとした．特徴ベクトルは，重心の x 座標 v1，重心の y 座標 v2，面積 v3，縦横比 v4 の 4 次元で表される．

最終的に，類似文書検索結果としてスライド類似度を降順にソートしたものをユーザに提示する．これによって，ユーザは領域特徴クエリで入力した部分領域のレイアウトに類似した順にランキングされたスライドを得ることができる．

5 おわりに

本稿では，社内文書の再発見型検索を実現するシステムとして，社内文書の視覚的な印象を分析・抽出し，ユーザの印象に近い文書の検索を可能とする視覚印象検索システムを提案した．

今後は，システム実装を行い，提案方式の評価を行う予定である．

参考文献

[1]
http://www.boxesandarrows.com/view/four_modes_of_seeking_information_and_how_to_design_for_them
[2]Volker Gaede, Oliver Günther, Multidimensional Access Methods, ACM Computing Surveys, Vol.30, No.2, pp.170-231, June 1998.