スタイル情報を考慮したPowerPointファイルの類似度算出方法

尾内　くらら

大学や会社のプレゼンテーションの場において, PowerPointは欠かせない存在となっている．また，企画書や製品説明パンフレット，会社案内といった幅広い用途において利用されている. このような状況に伴い, PowerPointに対する検索要求も多種多様化してきている．

現在のところ, PowerPointに対する検索は本文の内容に着目したテキストベースの検索が主流である．一方，PowerPointは通常の文書とは異なり，色・フォント・配置といったスタイル情報を主要な要素として有している．しかし，このようなスタイル情報に焦点を当てた検索方法の提案は数少ない.

そこで本研究では，色のみを考慮した類似度・色とフォントの２つの条件を考慮した類似度・色とフォントと配置のスタイル情報すべてを考慮した類似度といったように，各段階での類似度を算出できる類似度算出方法の提案を行う．これにより，利用者の多様な要求に合わせたPowerPointファイルの類似度検索を実現できると考える.

類似したスタイルをもつ PowerPoint ファイルの検索を実現するためには，PowerPoint ファイル間のスタイルの類似度を求める必要がある．本研究では，類似度算出の対象として，MS-Office で使用されている Office 文書の形式の OOXML(Office Open XML)を考える．PowerPointの標準のファイル保存形式は，Windows向けのPowerPoint 2007のリリースからOOXMLの形式となっている．OOXMLファイルは複数のパーツからなり，各パーツは複数の XML 文書から構成されている．この性質に基づいて，XML 文書の類似度検索を応用することにより，類似したスタイルをもつPowerPoint同士の検索を行う.

本論文で提案する類似度算出手法は，次の３つの特徴を持つ．1 つ目は，XML 文書を比較する際に, 要素名と属性名が一致したものの属性値を取り出し, 値に基づいた類似度を計算することである．2 つ目は，類似度の計算方法として，値の種類（RGB 値・配置 / 輝度，フォント名等）に応じて異なるものを用いることである．3 つ目は，「色」「配置」「フォント」の 3 つのスタイルにおいて，ユーザが自ら重み付けを指定して類似度算出できることである．以上 3 点の特徴を用いて，葉ノードの値の類似性を考慮した新たな類似度を定義することにより，類似したスタイルをもつ PowerPoint のより高精度な検索を行う．検索結果は類似度の高いものから順にランキングしたものを表示する.

評価実験では, 著者が作成したテストデータと Web 上に実際に存在する実データの PPTX ファイルに対して，適合率と再現率を評価した．その結果，多くの場合にデザインの類似した PPTX ファイルを取り出せること, そしてユーザが指定した重み付けを概ね反映した類似度のランキングが得られることを示した.

（指導教員　鈴木伸崇）