平成23年度 情報工学コース卒業研究報告要旨

渡邉 研究室	氏　　名	横 井　聡
卒 業 研 究 題 目	文書構造に基づいた論文ページの部分検索法	

今日では，記憶媒体が安価に入手できるため，従来紙媒体で保存されていた文書の電子化が進んでいる．その中でも，研究者は論文を文書画像として扱う機会が多い．しかし，論文画像を大量に保存すると，目的とする論文画像を発見することが困難になる．したがって，論文画像を検索するための手法が求められている．人間が閲覧したことのある論文画像を検索する場合は，ページの視覚的な特徴を利用できれば利便性が高い．そこで，本研究では図 1 に示すような人間が論文画像に対して記憶している視覚的な特徴から，論文画像を検索することを目的とする．人間が記憶している論文ページの特徴は部分的で曖昧であるため，記憶していない部分を補完し，目的とする文書画像の構造を推定することによって検索 (以下，部分検索) できる．

部分検索のために，図や表，テキストといった論文ページの中で視認性の高い要素の位置情報を反映できるインデックスを構築する必要がある．本研究では論文画像をページの構造に基づいて分類することによってインデックスを構築する．論文画像は，カラムや空白などによって区切られたエリアの集まりとして認識できるので，エリア間の相対的な位置関係としてページの構造を表現することができる．そこで，ページの構造と，エリア内に含まれる要素の縦方向の順序 (以下，エリアパターンと呼ぶ) に着目して論文画像を階層的に分類し，インデックスを構築する．これにより，論文ページに含まれる各要素の大まかな位置情報を表現することが可能である．

提案手法は，インデックス構築と部分検索の二つで構成される．インデックス構築では，論文画像を 1 ページずつ入力として与え，各要素領域を抽出し，構造とエリアパターンを特定することによって論文画像を分類する．部分検索では，目的とする論文ページのうち，ユーザが記憶している要素をクエリとし，残りの部分を補完することにより検索を実現する．ユーザは，ページを表す紙に見立てた領域上に，要素が存在する矩形領域を描画することにより，クエリを入力する．描かれた要素領域から目的とする論文ページの構造と各エリアのエリアパターンを推定することによって，クエリに適合する論文画像を検索する．

我々は論文画像を構造に基づいて分類し，部分検索を行うシステムを構築した．提案手法の有効性を検証するために 1003 枚の論文画像からインデックスを構築し，部分検索の精度を評価した．インデックスを構築した結果，論文画像は特定の構造に偏ることが分かった．また，提案手法による部分検索ではクエリに適合するする論文画像を 77 %の精度で検出でき，提案手法の有効性を確認した．

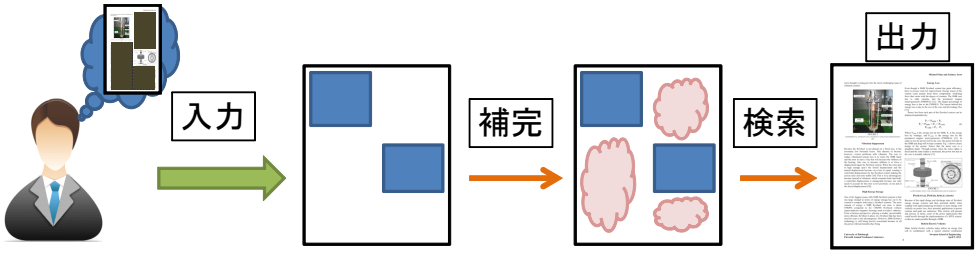

図 1 ページの視覚的特徴に基づいた部分検索