# 平成16年度 情報工学コース卒業研究報告要旨

| 吉川 研究室 | 氏　　名 | 杉木健二 |
|---|---|---|
| 卒業研究題目 | 学術情報のための文書の構造化に関する研究 | |

近年，インターネットの普及に伴い，学術情報の電子的な環境が整いつつある．電子図書館や国立情報学研究所をはじめとする多くの機関では，学術論文検索システムをWWW 上で提供しており，学術論文を対象とする研究もいつくか行われている．

現在，ほとんどの学術論文はPDF 形式やPS 形式のフォーマットで配布されている．これらは，配布や印刷のためのフォーマットであるので，データの抽出や加工，検索，再利用などは行いにくい．PDF 形式やPS 形式のファイルをテキスト形式やHTML 形式に変換するツールは多く存在するが，学術論文を利用するためにはこれらのファイルに人手で必要な情報を加えたり，部分的に抽出したりする必要がある．

そこで，本論文では，タイトル，著者，節，段落，引用，参考文献などの論文情報を含む，論文用のDTD を定義し，PDF 形式の論文をそのDTD に基づいた整形式XML データに自動変換するシステムを提案する．まず，PDF 形式の論文ファイルを中間ファイルに変換する．データ抽出プログラムを用いることで，一行のテキストごとに，フォント情報と位置情報を含むようなXML 形式の中間ファイルに変換する．次に，その中間ファイルから，論文のフォントサイズやページ上の位置などのレイアウト情報と，テキスト情報を利用して，抽出を行う．レイアウト情報から,「１ページ目でフォントサイズが最も大きくセンタリングされているテキストがタイトルである」などのルールを作成する．また,「概要」,「キーワード」,「はじめに」などの特定語や論文に固有の表現を表す正規表現を用いてテキスト情報からルールを作成する．

上述のルールを基にして抽出を行うが，テンプレートを用いた論文の場合はレイアウトがほとんど同じである．例えば，タイトルのフォントサイズや段落のページ左端からの位置などである．これらの論文に対しては，テンプレートに依存するレイアウト情報のルールも加える．その後，抽出した各行のテキストに対して該当するタグを付与した．引用や参考文献，著者，所属の場合は，これらの依存関係がわかるように，属性にラベルを付与する．一行ごとに抽出するので，概要などは，複数行に同一のタグが付与されている．これらの同一タグは結合させ，また，節タグは節全体にかかるように修正を行い，最終的なXML ファイルを出力する．

本システムの有効性を評価するために，本研究では国内の論文を対象とし，CD-ROM から収集することのできた情報系の論文を用いて実験を行った．その結果，本システムの有効性を確認した．これにより，現在流通している学術論文の自動変換により，XML 形式のデータとして利用できることを示した．