平成１５年度　情報工学コース卒業研究報告要旨

吉川 研究室	氏　　名	鈴木　祐介
卒業研究題目	構文情報を用いたWeb文書の自動分類	

インターネット上には多様なWeb文書が大量に存在している．この膨大な情報の中から必要な情報を効率良く探し出せることは，インターネットが日常的な情報源として利用されつつある近年において重要になっている．Web文書があらかじめ内容別に整理されていれば，分野や内容を絞った検索が可能になり，効率の良い検索の実現が期待できる．

本論文では，構文的情報を用いた文書分類手法を提案する．一般に，文書分類では，対象とする文書から分類の手がかりとなる特徴要素を抽出する．従来手法では，特徴要素として，単語の頻度といった文書の表層的な情報を利用することが多いが，文書の構文的情報も特徴要素として活用することにより，従来手法と比べ，より正確な分類を実現できる可能性がある．

本研究では，文書分類にベクトル空間モデルを使用する．ベクトル空間モデルでは，文書を多次元の特徴ベクトル空間として表現する．本研究では，文の依存構造に着目し，文書中から依存関係にある単語間の関係や文節間の関係を抽出して特徴要素に取り入れる．まず，文書を依存解析して単語間や文節間の依存関係を見つけ，単語間の関係からは依存関係にある単語対を，また，文節間の関係からは文節内で主辞となる単語対を抽出する．これらを特徴ベクトルの要素として加えることにより，単語情報と依存情報の双方を反映した特徴ベクトルを作成でき，構文情報を考慮した文書分類が実現できる．

本手法を用いてWeb文書の分類実験を行った．実験では，Yahoo!のサイトから14カテゴリに分類された学習用データ1002ページ，分類用データ99ページを用いた．依存情報を使用した本手法の正解率は51.5％であり，従来の単語情報のみを使用した場合と比べて3.0％の向上を示していることから，文書分類に構文情報を用いることの有効性を確認した．