文書のキートピックに基づくキーフレーズフィルタリング

田中　千尋

近年の情報技術の発達により，多くの文書情報が電子化されたことで，インターネットなどを通じて膨大な量の情報にアクセスできるようになった．膨大な量の情報の中から求める情報を探し出すためには，情報検索システムの補助が不可欠であり，盛んに研究が行われている．中でも，キーワード抽出の技術は対象の情報に自動的にタグをつけるなど，情報検索分野で活用されており，より精度の高い手法の提案が求められている．

本研究では，キーフレーズ抽出の精度を高めるために，用いる素性の調査が必要であることを指摘し，セクションを文書単位としたLDAの学習に基づいたキーフレーズフィルタリング手法を提案した．LDAでは複数のトピックを一つの文書が持つことを許容されているが，その中に文書を表すのにふさわしいキートピックがあるという仮定に基づいてキートピックと同じトピックに所属する単語を抽出することで，キーワードをどの程度フィルタできるかを検証した．

その結果，全単語の7%を残して，フィルタリングを行うことができることを示した．このことから，この手法は，キートピックの推定を行うことによってフィルタとしての利用も期待することが可能であるのみならず，素性としての利用も期待できる．

実験結果の考察から，キートピックを用いた提案手法では，適合率が対象の文書によってばらつきがあることが分かった．キートピックを用いることによるキーフレーズの絞り込みの効果の大きさに影響を与える文書の特徴について考察することは今後の課題といえる．

また，本研究では，キートピックを用いたキーフレーズフィルタリング手法の可能性を検証するために，キーフレーズが既知の文書を用いて，当該の単語に付与されたトピックをキートピックとして認定した．しかし，本来，キートピックはキーワードが不明な状態から推定する必要のある情報であり，キートピックの推定手法の提案が今後の課題である．

(指導教員　若林啓)