PageRank に基づく複数トピックをもつ文書に対応したキーワード抽出

池田　彩佳

インターネットの普及による電子文書の急増から，大規模文書群を扱うための技術が注目されている．なかでも文書の特徴を表す語を複数導出する自動キーワード抽出は，情報検索，文書要約，文書分類などに応用される基礎的な技術として文書管理の分野に大きく寄与する研究である．これまでに数多くのキーワード抽出手法が提案されているが，本研究では単語の共起構造を情報としてキーワードを得る手法を検討する．

単語の共起情報を用いて自動キーワード抽出を行う一般的な手法として Web ページのランキングアルゴリズムである PageRank を応用したものがある．しかし，この手法では文書内のあるひとつの単語の共起構造のまとまりからキーワードを抽出してしまう．文書は単一の話題で構成されるわけではなく，著者が主張を述べるために複数の話題や事実によって形成される．これは単語の共起から得られるまとまりで再現される．本研究ではこのまとまりをトピックと呼ぶ.

トピックの流れに即したキーワード抽出を実現している手法として KeyGraph がある．KeyGraph では文書中の著者の主張の流れを考慮するために，複数の共起しやすい頻出語の集まりをキーワードの核としてキーワードの選択をしている．

本研究では，PageRank に KeyGraph の構造を導入することで，より精度の高い複数トピックをもつ文書に対応したキーワード抽出法を提案する．本手法は DivRank と Topic-Sensitive PageRank によって構成する．提案手法は，(1)トピックのまとまりを得る，(2)トピックのまとまりからキーワードらしさを得る，という 2 段階で構成することにより，複数トピックをもつ文書に対応する．

評価はキーワード抽出の評価セットである NUS keyphrase corpus によって F 値を算出して行った．全評価文書における F 値の平均値は TextRank で 0.3687，KeyGraph で 0.3589，提案手法で 0.3717 であった. 提案手法において, トピックの核を DivRank 値が 0.05 以上の単語，キーワードを正解語の数に設定したときに最も F 値が高くなった．

提案手法は著者が新しいアルゴリズムや分析手法を提案している文書など，複数のトピックを組み合わせて記述された文書において有効であった．関連研究，手法，実験などがはっきりと章に分けられ構成されていたためと考えられる．一方，同じ単語が文書中一貫して用いられていた文書では TextRank が最も有効であった．

トピックどうしがあまり共起しあわない文書では複数トピックを考慮する手法が有効であることが確認できた．今後の課題として，文書の特徴ごとにトピックの分割度を調整するパラメータの設定，及びトピック分割の手法の検討が求められる．

（指導教員　若林啓）