論文内容の要旨

博士論文題目　Kernels for Structured Data in Natural Language Processing
(構造情報を利用したカーネルに基づく自然言語処理)

氏名　　鈴木 潤

(論文内容の要旨)

従来，自然言語処理の応用タスク (情報検索，テキスト分類等) では，出現単語の集合を特徴としてテキストを表現する，いわゆる bag-of-words モデルが主流であった．しかし，この手法は，本来テキストが持つ様々な言語情報を扱えないという問題がある．一方，近年では，テキストを「内容に基づいて分類」する，新しいテキスト分類タスク (sentiment classification) へと研究が移行しつつあり，テキストの持つ意味的・内容的な特徴を扱う方法が必須となった．

そこで，本論文では，テキストに内在する言語的な構造情報を利用する方法を提案し，これらの情報が「内容に基づく」テキスト分類タスクに対して有効であるかを検証する．また，本論文では，タスクに依存しない汎用的な枠組とするため，提案手法を全てカーネル関数で定義する．

はじめに，単語属性 N-gram を提案する．これは，単語が持つ品詞，辞書情報といった属性を組み合わせた N-gram であり，出現する様々な言語表現の利用が期待できる．質問分類タスクを用いて，有効性を検証するとともに，どのような素性が有効であったかを明らかにする．また，単語属性 N-gram が拡張 sequence カーネルとして定義できることを示す．

次に，HS-graph によるテキストの表現法と HS-graph カーネルを提案する．これはテキスト内の様々な言語的な構造情報を統合し，テキストの特徴として利用する方法である．提案手法により，従来扱うことが困難であった複雑な言語構造を，容易かつ効率的に扱うことを可能とした．また，効率的な計算方法を述べ，現実的な時間で提案手法が計算可能であることを示す．また，質問分類タスクを用いて，言語構造がタスクの性能向上に大きく寄与することを示す．

最後に，統計量に基づく素性選択法を導入した離散カーネルを提案する．離散構造を扱うカーネルには，全ての部分構造を均等に利用すると，学習時に過学習を起こす可能性が高いことが，これまでに報告されている．従来は，サイズの大きい部分構造（素性）を削除するか重みを下げることで，この問題に対応してきた．しかし，テキストの場合，特定の表現といった比較的サイズの大きい部分構造が有効な情報を持つ可能性が高いため，従来法では，これらの情報を効率的に扱えないという問題が残っていた．そこで，統計量による素性選択法を離散カーネルに適用することで，サイズに依存しない部分構造の利用法を述べる．また，マイニング手法を適用することで，高速なカーネル演算が可能であることも示す．文の主観・客観性を判定するタスクを用い，提案手法の有効性を検証するとともに，サイズの大きい部分構造が有効な特徴であることを明らかにする．

氏 名	鈴木 潤

（論文審査結果の要旨）

平成16年12月27日に開催した公聴会の結果を参考に平成17年2月17日に本博士論文の審査を行った．以下のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

鈴木 潤は，本博士論文において，文の構造情報を用いた種々のカーネル計算法を提案し，文書や文の分類を従来よりも高性能に行うことが可能であることを示した．本論文の貢献は次のようにまとめることができる．

1. 単語属性N-gramによるカーネルを提案した．これは，単語がもつ品詞や意味クラスなどの素性を複合的に組み合わせることのできるカーネルである．質問応答に用いられる質問文の分類タスクによる実験により，従来からの単語に基づく分類に比べて，有効であることを検証した．

2. 文の複雑な構造を利用する階層構造グラフカーネルを提案した．言語の文は単語が集まって文節や句を構成し，それらが更に統語構造を構成している．そのような構造を直接用いるカーネルの効率よい計算法を示した．質問分類タスクにより，構造情報を用いることによって，分類性能が大きく改善できることを示した．

3. 統計量を利用した素性選択法を提案した．複雑な構造を素性として用いるカーネルは，極めて多くの素性を扱うため，学習時の過学習を避けることができない．しかし，ある程度大きな構造が有効に働くような複雑な分類問題の場合には，素性の構造の大きさや頻度によって素性集合を足切りすることは得策ではない．本論文では，統計量を利用した素性選択により，分類タスクに有用な素性を抽出することにより，精度向上と高速化を達成する方法を示した．

このように，複雑な構造を利用することのできる種々の離散カーネルを提案し，文分類における有効性を示した本研究は，独創性が高く，しかも実用的であり，自然言語処理の分野において高い貢献があると評価する．

よって，本論文は，博士（工学）の学位論文として価値あるものと認める．