

(続紙 1)

京都大学	博士 (情報 学)	氏名	西 野 正 彬
論文題目	Numerical Optimization Methods based on Discrete Structure for Text Summarization and Relational Learning (文書要約と関係学習のための離散構造に基づいた数値最適化法)		
(論文内容の要旨)			
<p>本論文は自然言語処理における複数文書の要約と機械学習におけるオブジェクト間の関係の学習に対して、それぞれを実現するための数値最適化アルゴリズムを提案するものである。文書要約においては様々な要件を考慮する必要がある、関係学習においては複雑な関係を推定する必要があるため、既存の最適化アルゴリズムでは精度のよい解を効率的に求められないことが多い。本論文はこのような背景を念頭に置きながら具体的なアルゴリズムを提案している。提案するアルゴリズムは、タスクの実現のために解くべき数値最適化問題を小問題に分割し、それぞれの領域で扱うデータの持つ離散構造を活用するという設計方針に基づいており、結果として既存のアルゴリズムよりも高精度・高速な求解を実現している。</p> <p>第1章は序章であり、自然言語処理および機械学習における数値最適化問題と離散構造を用いた最適化問題の分割について述べている。続いて本論文の概要について、第3章で扱う確率論理プログラムのパラメータ推定というタスクおよび第4章で扱う隣接行列と実数行列の繰り返し乗算というタスクの関係学習における位置付けを述べながら示している。</p> <p>第2章では、複数入力文書の要約を自動的に生成する要約アルゴリズムについて述べている。提案するアルゴリズムは、組合せ最適化に基づく複数文書要約手法に関する先行研究において提案されていた関連性、冗長性、網羅性をそれぞれ表す3つの関数の和を同時に最大化することによって、既存手法よりもよい要約を求めている。一般に、複数の目的関数の和を最大化する最適化問題はNP困難であり厳密解を効率的に求めることはできない。しかし、提案するアルゴリズムは、ラグランジュ緩和法に基づき一致制約を目的関数に移した上で目的関数を分割し、さらに各関数の最適化問題を離散構造を利用して効率的に解くことにより、結果として高精度な近似解を高速に求めることを可能としている。実データを用いた実験により、既存の近似解法を用いるアルゴリズムと比較して、ROUGEスコアという指標において良い要約が得られたことを確認している。また、厳密解法と比較すると、目的関数の値の低下は6%程度に収まる一方で、厳密解法よりも平均で30倍以上高速に動作することを確認している。</p> <p>第3章では、数理論理学をもとにした論理プログラムを拡張した統計モデルである確率論理プログラムのパラメータを訓練例より推定するためのアルゴリズムを提案している。提案するアルゴリズムでは、負の対数尤度関数とペナルティ項の和を最小化する最適化問題を解く。その際、推定された確率論理プログラムにおける確率的な節の個数を削減できるようなペナルティ項を提案した。さらに論理プログラムの解釈のもつ離散的な構造に着目し、知識コンパイルという手法と射影勾配法を組み合わせたアルゴリズムを考案し、効率的に最適化問題を解くことに成功している。実データを用いた実験では、既存のEMアルゴリズムに基づくパラメータ推定手法と比較して同等の精度を達成しつつ、かつ確率的な節の数が少ない確率論理プログラムが推定できる</p>			

ことを示している。

第4章では、隣接行列と実数行列との繰り返し乗算を高速化するために、新たなデータ構造である隣接森(adjacency forest)を提案した上で、それを用いた行列乗算アルゴリズムについて述べている。隣接行列と実数行列との繰り返し乗算は、代表的な関係データ解析アルゴリズムにおいて用いられている。隣接森は、隣接行列を列方向に分割した上で、複数の行に重複している成分を集約することによりコンパクトに表現したデータ構造であり、隣接行列の乗算の高速化を達成する。隣接森の性質を理論的に解析し、また実データを用いた実験によって既存の疎行列表現を用いた時と比べて、隣接行列との繰り返し乗算が最大で300%高速化できたことを示している。

第5章では、本論文で与えた結果を概観し、今後の展望について述べて、結論としている。

注) 論文内容の要旨と論文審査の結果の要旨は1頁を38字×36行で作成し、合わせて、3,000字を標準とすること。

論文内容の要旨を英語で記入する場合は、400～1,100 wordsで作成し
審査結果の要旨は日本語500～2,000字程度で作成すること。

(続紙 2)

(論文審査の結果の要旨)

本論文は、自然言語処理における複数文書の要約と機械学習におけるオブジェクト間の関係の学習に対して、それらを実現する際に解くべき数値最適化問題を、既存のアルゴリズムよりも高精度で高速に求解するアルゴリズムを構成することを目的としている。これらの分野においては、タスクは複雑化しており数値最適化手法の応用が必須であるが、その際にはタスクの特性を利用した応用方法を設計しなければならない。本論文では、解くべき数値最適化問題を小問題に分割し、さらにタスクが扱うデータの持つ離散構造を活用するという一貫した方針を採用することにより、タスクに有効なアルゴリズムを提示している。主要な結果は以下の3つである。

1. 関連する複数の文書を要約した文書の生成を対象とし、先行研究で提案されていた、入力文と要約文の関連性、要約文の冗長性、要約文の網羅性をそれぞれ目的関数として表現し、それらの和を最大化するアルゴリズムを構成した。この問題はNP困難であるが、ラグランジュ緩和法に基づいて一致制約を目的関数に移した上で目的関数を分割し、各目的関数を離散構造を用いて効率的に最適化することにより、もとの目的関数の近似解を効率的に求めることを可能とした。さらに、実データを用いて、提案アルゴリズムが先行研究よりも良い要約を求めることができることを確認した。

2. 数理論理学に基づく論理プログラムの拡張となる統計モデルである確率論理プログラムのパラメータを訓練例より推定するタスクに対して、負の対数尤度関数とペナルティ項の和を最小化するアルゴリズムを構成した。ペナルティ項は、できるだけ確率的な節の個数が削減されるように設計するとともに、論理プログラムの解釈の持つ離散的構造を知識コンパイルという手法で圧縮することにより、効率的に最適化問題を解くことを可能とした。さらに実データを用いて、提案アルゴリズムは、先行研究において提案されているEM法に基づくアルゴリズムと同等の精度を達成し、かつ確率的な節の数が少ない確率論理プログラムを推定できることを確認した。

3. 代表的な関係データ解析アルゴリズムにおいて利用される隣接行列と実数行列との繰り返し乗算というタスクにおいて、隣接森(adjacency forest)という新たなデータ構造を利用した高速なアルゴリズムを構成した。隣接森を用いた行列乗算の計算を理論的に解析するとともに、実データを用いた実験により単純なアルゴリズムと比較して300%高速化できていることを確認した。

これらの研究成果はどれも全く新規な着想に基づいて行われている。本論文では、最適化問題を高精度かつ高速に解くアルゴリズムを設計する際に、問題分割と離散構造の利用という方針を貫いており、また、既存手法と同等以上の有用な解を出力あるいは高速に求解していることを実データを用いて確認しており、学術としての貢献度は高い。よって本論文は博士(情報学)の学位論文として価値のあるものと認める。

また、平成26年8月21日に実施した論文とそれに関連する内容についての口頭試問の結果、合格と認めた。

注) 論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。更に、試問の結果の要旨(例えば「平成 年 月 日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」)を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日： 年 月 日以降