審査の結果の要旨

氏名　趙　海燕

本論文は、A Compositional Approach to Mining Optimal Ranges (邦訳：最適区間探索の構成的手法に関する研究)と題し、最適区間探索の構成的手法に関する研究について述べたものである。データの収集方法の大幅な進歩と記憶装置の低価格化に伴って、膨大なデータを蓄積することができるようになり、こうして得られた膨大なデータ中に潜んでいる有意義な情報を抽出するデータマイニングの理論や技術の重要性が高まってきている。本論文では、データマイニングに対するアルゴリズムの効率化について、とくに最適区間探索の構成的手法を扱っている。期待する最適区間の抽出のための表現法として、本論文では簡便で一般性のある区間問合せ言語を提案し、その効率的な実現手法を示している。また、実現したシステムを用いて実際の POS データベースに対して適用した実験結果により、提案した手法の有効性を実証している。

第 1 章 Introduction では、本研究の背景と目的について述べている。膨大な蓄積データの分析に関する課題のなかで、とくにデータに潜んでいる区間情報を抽出することの意義を説明している。現実のデータに対しては、一つの区間だけではなく、複数の区間に対する最適区間を求めることも必要とされ、従来のアルゴリズムでは解決されていない問題点を指摘している。本論文では、このような最適区間問題に着目し、「あるデータが与えられたとき、性質を満たす最適区間を求める」一般的なアルゴリズムの構成法を対象としていることを述べている。

第 2 章 Range querying language では、最適区間を探索する実例から、最適区間の表現法とそのための言語の要素をまとめて説明し、RQL(Range Query Language)という、関係データベースの操作言語 SQL(Structured Query Language)に似た表現の言語を定義している。

第 3 章 System overview では、提案した言語 RQL を実現するシステムを概観し、そのために解決すべき課題を述べて、本研究の具体的な問題を明確に示している。

第 4 章 Compilation では、最適区間を抽出するための前処理について述べている。この前処理は対象とするデータと言語 RQL による表現の両方に対して行なっている。データの前処理には、すでに知られている準線形時間のアルゴリズムを用いることができるが、言語の表現を効率的に実現するには、条件を表現する述語のあらたな標準化(normalization)が必要とされる。本章では最適区間に関する述語の標準化手法を示して、そのアルゴリズムを述べている。ここでは、区間に関する性質を表現するために、求められる区間の最左要素と最右要素の間に成り立つ順序関係による述語に着目している。

第 5 章 One dimensional range problem では、一次元の最適区間問題の実現手法

を述べている。データマイニングにおける一次元の最適区間問題は、既知の非負最長部分列の問題に帰着されることを述べ、これが第 4 章で扱った標準形になっていることを示している。標準形の述語に現われる関係式が 1 個の場合には、すでに線形時間アルゴリズムが知られているが、複数の関係式によって記述される述語に対して効率のよいアルゴリズムはこれまでに知られていない。この場合には、第7章で述べている極小属性値を持つ多次元探索木を用いて解決できると主張している。

第 6 章 Multiple range problem では、多次元の最適区間問題の実現手法について述べている。一次元の最適区間の解決策を基礎として、二次元の最適区間問題に対するアルゴリズムを提案し、多次元への拡張を論じている。

第 7 章 Multi-dimensional searching trees with minimum attribute では、最適区間の探索に対する基礎としての最小属性値を持つ多次元探索木を扱っている。多次元空間における極小性を判定するアルゴリズムが、述語が複数の関係式の論理積で表現された場合の最長部分列問題の解決に有効であることを述べ、これまでには得られていなかった効率的な検査のためのデータ構造を提案している。ここで提案した最小属性値を持つ多次元探索木(k-d-m tree)により、k 次元空間における極小値であるかどうかを O (log ^(k-1) n) 時間で判定できるアルゴリズムを示している。

第 8 章 System for mining optimal ranges では、実現したシステムについて、その概要を述べ、ユーザインターフェース、マイニング条件の指定方法、結果画面の表示法等を示した上で、実際のデータを用いた実験結果を示してシステムの性能を評価し、アルゴリズムの有効性を示している。

以上、これを要するに、本研究は、大量のデータから希望する条件を満たす最適区間を抽出するという問題に対して、条件の表現法を提案するとともに、その条件に対する効率的なアルゴリズムの一般的な構成法を与えたものである。また、その方法によってプロトタイプシステムを構築して、実際に現実のデータに対して実用性を確認している。この成果は情報工学上貢献するところが多大である。よって本論文は博士(工学)の論文として合格と認められる。