# 論文の内容の要旨

論文題目　A Compositional Approach to Mining Optimal Ranges
(和訳　最適区間探索の構成的手法に関する研究)

氏名　趙 海燕

本論文は、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、記憶装置の劇的な低価格化に伴って、非常に膨大なデータを蓄積することができるようになった。得られたデータをもとに、未来の経営戦略を立てることが、効果的な経営に重要となってきている。このように、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への需要がここ数年大きくなってきている。これに伴い、データマイニングに対するアルゴリズムの効率化が重要となってきている。本論文には、特に最適区間探索の構成的手法に関する研究を行う。最適区間を効率的に抽出するため、本論文は使いやすい区間クエリー言語を提案して、効率的な実現手法を示す。実装したシステムを喫茶店のPOSデータベースのような現実データを用いてテストした。実験の結果から、本論文提案した手法は高速かつ実用的であることが検証された。

具体的に、本論文の枠組みと各章の内容を以下に述べる。

第1章＜Introduction＞では、本論文において行う研究の背景と目的について述べている。

身の回りをみると、いたるところでデータが収集されていることに気がつく。デパートやスーパーマーケット、コンビニエンスストアではＰＯＳ端末と呼ばれる機械で、バーコードを読み取って、商品の在庫や顧客層の情報を蓄えている。このようにして蓄積された膨大なデータの中に, 大量な有益な情報が潜んでいる。そこでは、この膨大な蓄積データをどのように分析すればよいのかということが課題となっている。そのような膨大なデータの中に潜んでいる規則性を抽出して、有用な情報を見つけ出すことを「データマイニング (data mining)」という。

本論文の目的はデータに潜んでいる区間情報を抽出することである。例えば、スーパーで売上げた商品量が時刻と共に記録されているものとする。例えば、以下の表

| ... | milk | bread | time |
|---|---|---|---|
| ... | *500* | *360* | 15 : 30 |
| ... | *680* | *200* | 15 : 42 |
| ... | *360* | *600* | 15 : 55 |
| ... | *400* | *150* | 16 : 10 |
| ... | *120* | *435* | 16 : 40 |
| ... | ... | ... | ... |

といった具合である。各行はある時刻にされた買物の状況を書かれている。例えば、1 行目のデータは 15 時 30 分になされ、牛乳とパンの買った量である。営業収益を上げるため、どの時間帯で牛乳が売れるのかという情報を得ることは営業者に対して有用である。そうすると、この時間帯で大量な牛乳を用意することできる。例えば、牛乳の売上げ量は平均的に 400 であり、最小も 100 以上の時間帯を求める。もちろん、この時間帯はできるだけ長いほうがもうかるので、求めるのは前述の条件を満たす最長の時間帯、即ち、最適の時間区間である。

現実のデータに対して、一つの区間だけでなく、複数の区間に対する最適区間を求めることが必要となってきている。これは多次元の最適区間である。

本論文には、上に述べた最適区間（一次元と多次元と共に）問題を目指して研究を行う。最適区間問題とは、簡潔に言えば、「あるデータ $x$ が与えられたとき、性質 $p$ を満たす最適の部分を求める」という問題である。本論文で解決できる性質 $p$ は図 1 に示された述語である。

第 2 章＜Range querying language＞では、本論文が提案した区間クエリー言語を紹介する。

最適区間を探索する実例から、最適区間の特徴と使いやすい言語の要素をまとめる。これらを目安として、本論文は $\mathcal{RQL}$ と呼ばれる SQL 風の言語を提案する。実例を通じてその文法と意味を説明する。特に中心として **FIND** ステートメントと述語（図 1 を参照) という区間が満たされるべき性質 (range property) を論ずる。

第 3 章＜System overview＞では、提案された言語を実装するシステムを概観する。システムの全体像を描く上、この言語を実現するときの核心問題をハイライトする。

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る。この前処理はデータと言語両方に対して行われる。

| | | | |
|---|---|---|---|
| $p$ | ::= | (sum *ColumnName*) $\otimes$ c | Sum Property |
| | \| | (avg *ColumnName*) $\otimes$ c | Average Property |
| | \| | (count *ColumnName*) $\otimes$ c | Count Property |
| | \| | (min *ColumnName*) $\otimes$ c | Min Property |
| | \| | (max *ColumnName*) $\otimes$ c | Max Property |
| | \| | $p \wedge p$ | Conjunction |
| | \| | $p \vee p$ | Disjunction |
| | \| | $not\,(p)$ | Negation |

図 1: 区間性質を指定する述語

最適区間を求めるため、元データは区間に関する属性によって分類されなければならない。この膨大なデータを分類するには、実用的な方法が必要となっている。これで、ある準線形の bucketing アルゴリズムを用いて元データを bucketing する方法を紹介する。 これによって、仮にデータを属性 $A$ によって分類すると、bucketing した後のデータには、$A$ の値域は

$$B_1, B_2, \ldots, B_M$$

$$(B_i = [x_i, y_i] \text{ かつ } x_i \leq y_i < x_{i+1})$$

のような連続 $n$-区間になる。

そして、言語を効率的に実現するため、それを簡約 (normalization) することが必要であるので、本章で最適区間に関する述語の簡約を明示する。簡約した後の述語は以下のような選言標準形をしている。

$$\begin{array}{rll} p = & p_{11} \wedge p_{12} \wedge \ldots \wedge p_{1k_1} & \vee \\ & p_{21} \wedge p_{22} \wedge \ldots \wedge p_{2k_2} & \vee \\ & \ldots\ldots & \vee \\ & p_{m1} \wedge p_{m2} \wedge \ldots \wedge p_{mk_m} & \end{array}$$

そのうち、各 $p_{ij}$ は以下の特別な性質を持っている形をしている：

$$\mathrm{f}\ (head) \otimes \mathrm{g}\ (last),$$

$\otimes$ は $\leq$ のような全順序の関係である。

第 5 章＜One dimensional range problem＞では、一次元の最適区間問題の実現手法について説明する。

データマイニングにおける一次元の最適区間問題は実際に最長部分列の問題である。というのは、「ある列 $x$ が与えられたときに、$x$ の性質 $p$ を満たす最適の部分列を求める」という問題である。簡単な例として、列

$$x = [x_0, x_1, \ldots, x_{n-1}]$$

が与えられたとき、$x$ の連続する部分列 (*segment*) $[x_i, x_{i+1}, \ldots, x_j]$ のうち、性質 $p$ を満たすようなもののうち、最長のものを求めることである。 ここで $p$ は言語 $\mathcal{RQL}$ で定義されて、第 4 章で得た簡約形である。そのうち、基となる問題は、両端要素が大小関係を満たす最長部分列の問題である：

> 数値列 $x = [x_0, x_1, \ldots, x_{n-1}]$ が与えられた時、$x$ の連続する部分列 $[x_l, x_{l+1}, \ldots, x_r]$ の左端要素 $x_l$ と右端要素 $x_r$ の間に、$x_l \otimes x_r$ という関係が成り立つようなもののうちで、最長のものを求める。

本章ではこのような基本的な最長部分列問題に対してある効率のよい（線形時間）アルゴリズムを提案し、その上で、最小属性値を持つ多次元探索木を用いて基本的な問題を組み合わせた複雑な問題（述語が論理積である場合）を解決することについて述べる。

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る。

一次元の最適区間の解決策を基づいて、二次元の最適区間問題に対するあるアルゴリズムを提案した上、多次元への拡張することを論ずる。

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紹介する。

多次元空間における最小値を検索することは実用と理論両面で有益なことである。これは元々計算幾何分野の問題であるが、本論文に述語が論理積である場合の最長部分列の解決することにも大変役に立つ。一方、最小値を検索する効率は基礎をなすデータ構造に依存している。本章では最小属性値を持つ多次元探索木（$k$-d-m 木に略す）という新しいデータ構造を提案し、また $O(\log^{(k-1)} n)$ 時間で（$k$ は次元の数）ある所与の多次元値が最小値かどうかを判断できるアルゴリズムを実現する。

第 8 章＜System for mining optimal ranges＞では、実装したシステムの枠組みを明示する。ユーザの視点から、システムのインターフェース、マイニング条件の指定方法、及び結果画面を展示する。また、現実のデータを用いてシステムをテストすることでシステムの性能を評価する。