学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　陸　可鏡

審査担当者　主　査　教　授　工藤 峰一
副　査　教　授　今井 英幸
副　査　教　授　田中 章

学位論文題名

Query-Aware Locality Sensitive Hashing for Similarity Search Problems
（類似検索問題に対するクエリ・アウェア局所鋭敏ハッシング）

近年のウェブ前提社会において「類似 (対象) 検索」は欠くことのできない一つの重要な技術となっており、その高速化は技術面だけでなく社会的にも大きなインパクトを与える。

本研究は、距離を基準とした「最近隣 (対象) 探索」(以降、NNS) と類似度を基準とした「最大内積 (対象) 探索」(以降、MIPS) の二つの「類似 (対象) 検索問題」を扱っている。いずれの問題の解も、サンプル数 n と次元数 d の積の計算量で全数探索することで求められる。しかし、n がギガを容易に超すような巨大な問題をしばしば扱う現代においてはわずかであっても計算量を n に関して劣線形に抑えることが重要となる。一方残念ながら、d が大きい高次元においては、NNS 問題や MIPS 問題を解くには全数探索が最も高速であることが知られている。結果として、準最適、つまり、ほぼ正解に近いサンプルを見つけることでよしとして、近似 NNS 問題 (ANNS) あるいは近似 MIPS 問題 (AMIPS) を n の劣線形時間で解く研究がこの分野では主流になっている。本研究においては、局所性鋭敏型ハッシュ (以降、LSH) と呼ばれる確率的ハッシュ法を中心技術として据え、個々のデータの特性を最大限引き出すことでこれまでの手法を計算量あるいは精度で改善している。

本研究では、NNS に対して二つ、MIPS に対して一つの手法を提案しているが、いずれの手法も精度保障がある点でこれまでの多くの高速化手法と比べ有利である。つまり、所与の近似精度を達成する確率を十分高くできる、いわゆる Probably Approximately Correct アルゴリズムになっている。また、大規模問題を想定しているためデータはメモリ上でなく外部記憶装置におくことを前提としている。

最初の ANNS に対しては VHP と R2LSH という二つの手法を提案している。VHP では 2 次元同心円状にハッシュバケットを設けることで、R2LSH ではハッシュバケットの衝突回数に加えて質問サンプルとの距離を考慮することで、これまでの手法より効果的に不必要な距離計算を省いている。また、query-aware LSH と呼ぶ、質問サンプルをハッシュ後に中央 (原点) に位置させられる方式を採用することで探索効率を高めている。従来手法の探索計算量が O(n log n+d) であったのに対して、提案の 2 手法は O(nm+d)(ここで m はハッシュ関数の数) と n が大きい問題において有利になっている。二つの提案法の比較においては、探索に必要な索引サイズにおいては VHP が、探索速度においては R2LSH が有利となっている。

二番目の (A)MIPS に対しては AdaLSH という手法を提案している。内積がノルムとコサイン角度の積であることから、ノルムを先に量子化し、より大きなノルムに対してより細かな量子幅

でコサイン角度を調べる 2 段階量子化方式を採用することで探索効率を高めている。この場合にも query-aware LSH を利用している。従来手法の探索計算量が O(n log n +d) に対して、提案法は O(nm+d) と、こちらも ANNS のときと同様な改良となっている。

これら三つの提案手法に共通する利点は、出力に確率的保障があることである。所与の近似誤差 (ANNS では真の NN と質問サンプルの距離の何倍までを解として許すか、AMIPS においては、どの程度までなら達成可能な最大内積の値が下がってよいか) と、所与の誤り率 (近似解を出力できない確率の上界) に対して必要なハッシュ関数の数 m や終端条件を導出している。このため、精度保障を持たない手法と比べて安心して使用できる。特に、近似誤差を零、つまり、真の解を求めるようにも指示できる点が新しい。実際的有効性を多数の大規模データベースを利用した実験で確認している。

本論文の貢献は以下にまとめられる。

1. 近似最近隣探索において、従来の手法に比べ計算量的に有利な二つの手法を提案し、その効果を理論および実験により確認した。
2. 近似最大内積 (対象) 探索において、従来の手法に比べ計算量的に有利な一つの手法を提案し、その効果を理論および実験により確認した。
3. 共に、(確率的) 精度保障があり、近似ではなく正確な解を求めるようにも設定することもできる点でこれまでの高速化手法より利用価値が高い。

これを要するに筆者は、類似探索問題において、数理科学的方法論に基づいて探索問題の性質を詳しく解析することにより、精度や速度においてこれまでの探索方式を上回るだけでなく精度保障がある探索手法を開発した。この成果はデータベースおよびパターン認識の分野に貢献すること大なるものがある。よって、著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。