論文審査の結果の要旨

氏名　岡　　敏生

本論文は「近似最近傍点探索と分散アプリケーションへの適用に関する研究」と題し，類似検索における基本処理の一つである近傍点探索を取り上げ，その効率的な実現方法および適用方法について論じている．近年，世界中のデジタルデータの量は膨大なものとなっているが，その中でわれわれ一人ひとりにとって有益なデータはごく一部であり，それを見つけ出すという作業が重要となる．こうした問題意識のもと，本論文ではデータセット全体からあるデータに類似したデータを探すという処理を扱っている．具体的にはあるデータの近傍に存在するデータを高速に見つけ出すアルゴリズム，およびその処理を分散的に処理するための負荷分散アルゴリズムについて検討している．またそれらの適用分野の一つとして協調フィルタリングについて着目し，適用する際の検討事項についても考察している．

第 1 章は序論であり，研究の全体的背景と各研究の研究内容の概要について触れ，本論文の構成について述べている．

第 2 章「高次元 Lp 空間における近似最近傍点探索アルゴリズム:sc-LSH」では，Lp 空間($0 < p \leqq 2$)，特にユークリッド空間における効率的な近似最近傍点探索アルゴリズムについて検討している．高次元空間における最近傍点探索では「次元の呪い」と呼ばれる現象によって探索の処理性能が落ち，総当たりによる線形探索と同程度かそれ以下の性能しか得られないことが広く知られている．そのため近年では近似解を高速に取得しようという研究が活発になされている．本章ではまず近似最近傍点探索問題の定義，既存研究等について言及したのち，sc-LSH というアルゴリズムを示している．sc-LSH ではランダム近傍点探索テーブルという概念を導入して高速化を図っている．具体的には，ランダムプロジェクションと呼ばれる手法でデータセットの次元を削減したのち，予め作成した探索テーブルにもとづいてデータ格納位置を決定している．これによって近傍点がある確率でハッシュテーブルにおける同じバケットに格納されることになり，探索の効率化が図られる．

第 3 章「分散環境における近似最近傍点探索」では，連想配列型データの局所性を考慮した負荷分散手法について検討している．連想配列は(key, value)ペアの形式で表現できる抽象型データ型であり，第 2 章の近似最近傍点探索を含め広範なアプリケーションに利用されるデータ型である．連想配列を分散的に実現する場合，分散ハッシュテーブルというアルゴリズムが利用されることが多いが，このアルゴリズムでは特定のキーに負荷が集中したときに負荷の均一化がなされない，キーの局所性が維持されないという

問題があった．そこで本論文ではブロックリダイレクションという方法を導入することで，キーの局所性の維持と負荷の均一化という問題を同時に解決している．これによりシステムの平均負荷がほぼ一定であり予め定数倍の範囲で事前に分かっている場合，任意のデータ分布に対して最も負荷が重いサーバの負荷は高い確率で平均負荷の$\Theta(\log\log N/\log d)$倍になるという負荷分散特性が得られている．また平均負荷について事前知識を持たない場合にも，良好な負荷分散が実現できることをシミュレーションによって検証している．

第4章「協調フィルタリングプラットフォーム Vineyard」では，先の第2章および第3章のアルゴリズムの協調フィルタリングプラットフォームへの適用について検討している．協調フィルタリングとはユーザ同士がアイテムに対する評価情報を共有することで，互いの情報収集を効率化するという概念を指している．協調フィルタリングでは評価情報に関してユーザ間の類似性やアイテム間の類似性を利用することが多いため，近似最近傍点探索を適用することが可能である．ここでは近似最近傍点探索や負荷分散アルゴリズムを適用するうえで，具体的にどのように実現すればよいかについて検討している．

第5章は結論であり，本論文の成果をまとめるとともに，規模拡張性のある類似検索システムを実現する上で残された課題について述べている．

以上，これを要するに，本論文は規模拡張性のある近似最近傍点探索システムの構築方法とそのアプリケーションについて考察したものである．大規模データセットから効率的に必要なデータを選び出すという処理は高度に情報化された現代社会において不可欠なものであり，情報通信工学上寄与するところ少なくない．

したがって、博士（　科学　）の学位を授与できると認める。