論文内容の要旨

論文題目　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

氏名　　岡　敏生

デバイス技術や計算機科学等の進展にともない，デジタルに処理することのできるデータの量は年々急激な増加を見せている．この膨大な量のデータを人がすべて一つずつ吟味して利用するということは現実的な話ではないため，自動的な仕組みによってデータを処理し，利用者にとって自然な形で情報なりサービスを提供することが求められる．このようなことを実現するうえで必要となる機能は当然目的によって異なり多種多様ではあるが，なかには広範な用途に利用可能な基本的機能というものが存在する．近傍点探索アルゴリズムが提供する機能もそのようなものの一つで，類似検索が必要とされるさまざまな用途に利用することができる．

本研究ではそのようなことをふまえて，近似最近傍点探索を行うための効率的な探索アルゴリズムとそれを分散処理するための技術，およびそれらの適用アプリケーションについて研究した．ここでいう近似最近傍点探索とはあるメトリック空間上のデータセットが与えられたとき，指定したデータに最も近いものを近似的に探し出すという問題をさしている．本研究ではメトリックとして特に Lp ノルム($p \leqq 2$)に着目した．Lp 空間はユークリッド空間を含めそれ自体でも応用の多い空間であるが，埋め込み(embedding)という処理によって他の空間から小さな歪み(distortion)で Lp 空間にマッピングできるケースがある

ため，Lp 空間における近傍点探索は実用上非常に重要である．ここで扱う近似最近傍点探索アルゴリズムはデータセットが大きい場合，基本的に大きな記憶領域（メモリもしくはストレージ）と多くの計算処理を要するため，対象データが大きい場合や処理すべきクエリ(query)数が多い場合，処理の分散化は避けて通れない．そこで本研究では近似最近傍点探索アルゴリズムの効率化を図るだけでなく，分散処理についても検討を行った．具体的には連想配列で表現されるデータを処理する際，局所性を考慮したうえで負荷が均一になるような負荷分散アルゴリズムを実現している．また近似最近傍点探索アルゴリズムを用いた具体的なアプリケーション例として協調フィルタリングに着目し，どのように適用すればいいかについても考察を行った．

本研究で扱った内容は，近似最近傍点探索アルゴリズム，局所性を考慮した連想配列型データの負荷分散，協調フィルタリングの 3 つの項目から構成される．以下では各項目について概説する．

近似最近傍点探索アルゴリズム

最近傍点探索とはあるメトリック空間においてクエリ点に最も近い点を求めるという問題である．一般に近傍点探索アルゴリズムはあるオブジェクトに類似するものを検索する際に利用できるため，文書検索，協調フィルタリングをはじめ，パターン認識，データマイニング，ゲノム解析等の広範なアプリケーションに用いられる技術である．大きなデータセットを扱う場合には近傍点探索を効率的に行う必要があるが，対象が高次元データであるとき「次元の呪い」と呼ばれる現象により効率的な探索が困難になることが知られている．このような状況を打開するため高速な近似アルゴリズムを見つけようという研究が活発になされている．

本研究では d 次元 Lp 空間において効率的に近似最近傍点探索をおこなう手法について検討した．具体的には，予め k 次元($k < d$)空間内に超球をランダムに配置しておく．そしてデータの次元をランダムプロジェクションによって d 次元から k 次元に削減し，超球がある空間にマッピングする．データを格納するさいには同じ超球内に存在するデータが同じバケットに格納されるようにハッシュテーブルに登録してゆく．近傍点の探索はクエリに該当するバケットをルックアップすることで行う．元の空間で近傍に存在する点同士は，そうでないものと比べて同じバケット内に格納される確率が高いことから，一部のバケットを探索するだけで近傍点探索を取得することができる．

この一連の処理では各データ点を包含する超球を特定する処理が鍵になってくるが，本

研究ではこの処理を行うためにランダム近傍点テーブルというものを導入した．予めこのテーブルを準備しておくことで，各データ点を包含する超球のルックアップを高速に行うことが可能になる．

局所性を考慮した連想配列型データの負荷分散

つづいて先ほどの処理を分散処理するための技術について述べる．先ほどの処理ではハッシュテーブルを用いて連想配列にデータの格納を行っていたため，ここでは連想配列を分散的に実現する方法について扱う．連想配列とは(key1, value1), (key2, value2), (key3, value3), ...の形式で表現される抽象データ型のことで，広範な用途に利用できるデータ型である．ここではキーの重複を許すものとして扱う．本研究で実現する負荷分散は以下のような要求を満たす負荷分散である．

要求：

1，各ホストに格納される(key, value)ペアの数をほぼ均等に分配したい
2，キーを与えられたときに高速に値にアクセスしたい
3，特定のキーに該当する値が大量に存在する場合にはそれらをまとめて処理したい
4，あるメトリックによって近隣に存在するデータを同時にアクセスするときに効率的にアクセスしたい

3は4の特殊な形態であることに注意されたい．先ほどの近似最近傍点探索を分散処理する際には必ずしも4の条件を満たす必要はないが，1, 2, 4を満たすことでより広範なアプリケーションの負荷分散に役立つため本研究ではこれらの要件を満たすような負荷分散を実現する．

分散環境における連想配列の代表的な実装方法としては分散ハッシュテーブルを利用する方法があるが，通常の分散ハッシュテーブルではアイテム毎にキーにもとづいて格納サーバをランダム化するため1, 4の要件を満たすことができない．そこで本研究ではブロックリダイレクションという手法を導入して，上記の要件を満たす負荷分散を実現する．つまりアイテム毎に格納サーバをランダム化するのではなく，キー同士が近隣に存在するデータをまとめて均一サイズのブロックを形成し，ブロック単位で格納サーバをランダム化するということを行う．これによって1，4の要件を満たすことが可能になる．また本研究ではブロックを配置する際，単純にランダム化するのではなく，ビン・ボール過程におけ

る"the power of two choices"というパラダイムを適用している．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($d > 2$)になるという負荷分散特性が得られる．

協調フィルタリング

デジタルに利用できる情報の増加に伴い，その中からユーザに求めるものにもっとも適したものを人手で探すことが困難になってきている．そのような状況を解決するため，ユーザ同士のプロファイル情報を利用することで情報収集を効率化しようという動きがある．このようなものを総称して協調フィルタリングと呼ぶが，代表的な協調フィルタリングアルゴリズムではユーザ間のプリファレンスの相関やアイテム間の特徴量の類似度にもとづいてアイテムのランキングを生成している．

このようなシステムで重要となるのはデータセットサイズが大きくなったときにどのように対処するかである．ユーザ数やアイテム数が多い方が，対象ユーザに興味の近いユーザがいる可能性やユーザの要望に適したアイテムが存在する可能性は高くなるが，その一方で処理コストの増大を招く．

本研究では次のような方法でこの問題を解決する．処理コストの増大に関して，前述の分散近似最近傍点探索システムを用いることによって対処する．協調フィルタリングにおいて計算量が必要となるのは，プリファレンスの近いユーザや特徴量の近いアイテムを探す処理であり，この処理部分を近似アルゴリズムで代用することによって計算量を削減することが可能となる．加えてアイテムのメタデータを利用することで対象アイテムを限定することについても検討を行う．