公開

研 究 終 了 報 告 書

「超高速 IoT ビッグデータ解析のための分散アルゴリズム基盤」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 天方 大地

1. 研究のねらい

IoT デバイスによって世の中の事象をセンシングしたデータを大量に生成し、それら（IoT ビッグデータ）を保存・共有することは今日では当たり前である。一方、大量にデータがあるために、アプリケーションが求めるデータの処理・解析に要する時間は増加傾向にあり、ビッグデータを高速に処理する技術は増々重要となっている。これまでに様々な解析方法、およびそれらを対象とした高速アルゴリズムが提案されているが、IoT ビッグデータを高速に解析する技術は十分に開発・実装されていない。これは、以下の特徴が要因の一つである。

(i) 異種性：大規模な環境では様々なデバイスを用いて様々なドメインをセンシングする。センシング空間での事象解析のために各ドメインを評価する必要があるが、それぞれのドメインに必要なデータ空間（距離または類似度の指標）は自明ではなく、様々なデータ空間でテストできることが望ましい。また、各ドメインは独立しているため、データ型もドメイン間で統一されていない。そのため、入力を任意のデータおよび距離指標とするアルゴリズムが必要となる。

(ii) 大量のノイズの存在：IoT データにはノイズ・アウトライア（外れ値）が含まれることが一般的であり、無駄なデータ量の増加および計算コストの増加の原因となる。そのため、ノイズを高速に検出・除去するアルゴリズムが必要となる。

本研究では、特定の解析方法を提案するのではなく、多くの解析ツール・システムで用いられるプリミティブオペレータの処理を超高速に実行する並列分散アルゴリズムを設計する。具体的には、既存の技術に比べて数百倍の高速化を目指す。数百倍の高速化の恩恵の代表例として、既存の解析処理をインタラクティブに実行できるようになることや、これまで扱えなかったサイズのデータ量を処理できることが挙げられる。また、ビッグデータ解析は複数のプリミティブオペレータを組み合わせて行うものがほとんどと言えるため、それらを高速化する効果は非常に大きい。

分散処理システムでは、あるアルゴリズムは最大でもワーカ（計算機や CPU コア）の数に比例する程度でしか早くならない（実践的にはワーカ数に比例するほどの高速化は起こらない）。数百倍早くするためには、数百のワーカが必要となるが、これは管理的にも経済的にも現実的でなく、1 台の計算機を想定した提案システムの方が消費電力的にも環境に優しい。

2. 研究成果

(1) 概要

本プロジェクトでは、以下の問題に主に取り組んだ。

(i) 静的データに対する密度ベースクラスタリング問題。Science で発表された Density-Peaks Clustering の高速化に取り組んだ。このクラスタリングは様々な分野で利用されており、効果的なクラスタを提供するが、計算時間に問題があり、データ

数に対して自乗時間のコストを要した。この課題に対して、自乗未満時間で実行できる条件を示した。

（ii） 動的データに対する密度ベースクラスタリング問題。上に対して、データが一つ追加または削除された時、メトリック空間においてクラスタの正確な更新にかかる時間がΩ(n)時間であることを示した。つまり、正確な更新は低速であるため、近似アルゴリズムを設計した。

（iii） 距離に基づくアウトライア検出問題。距離に基づくアウトライア検出は様々な分野で応用されているが、データ数の自乗時間かかることが欠点であった。そこで、近接グラフを使った新たなアプローチを提案し、O((t+f)n)時間で動く（t はアウトライア数、f はフィルタリングされなかったインライア数、n はデータ数）ことを示した。

（iv） 高次元空間における近似 k 最近傍検索問題。埋め込みベクトルの普及に連れて、高次元空間での最近傍検索の需要が高まっている。この研究では、既存のアプローチを組み合わせて、所謂良いとこ取りを実現したアルゴリズムを設計した。また、マルチスレッディングにより高精度となる確率が高くなることを理論的に示した。

（v） k 最近傍距離推定問題。密度推定のように、k 最近傍までの距離のみ知りたい場合、検索は必要ないため、距離が推定できればよい。これを高速・高精度で実行できる機械学習モデルを設計した。

（vi） 逆最大内積探索問題。あるアイテムがどのくらいのユーザの好みにマッチするかは、市場調査等の指標となる。EC サイトには大量のユーザ・アイテムが存在するため、この操作を効率的に実行することは自明ではないところ、これを高速に実行できるアルゴリズムを設計した。

（vii） 多様性を考慮した最大内積探索問題。推薦システムでユーザの好みに合うアイテムを提供することは重要であるが、推薦リスト内のアイテムに偏りがある場合、新たな発見・気づきが失われる。この課題を解決するため、推薦リストを多様化する問題を定式化した。さらに、無駄なデータアクセスを削減して高速に推薦リストを検索するアルゴリズムを設計した。

（2）詳細

(i)「静的データに対する密度ベースクラスタリング問題」

この研究では、2015 年に Science で発表された Density-Peaks Clustering の高速化に取り組んだ。このクラスタリングは、各データを中心とする一定の領域内に存在するデータの数を local density と定義し、自身よりも local density が大きいデータの中で最も近いものを dependent point と定義する。dependent point との距離が大きいデータは密度がピークに達していると想定し、このデータをクラスタセンタとしてクラスタが形成される。このクラスタリングは様々な分野で利用されており、効果的なクラスタリング結果を提供することが報告されている。一方、各データの local density および dependent point の計算には O(n)時間かかるため、クラスタリングの時間は O(n^2)となってしまい、大量データにスケールしないという問題があった。

そこで、大量データでも短時間でクラスタリング結果を返すアルゴリズムを設計した。1つ目の提案アルゴリズムは厳密解を返す。このアルゴリズムは kd 木と呼ばれる木構造のインデックス

を利用することで高速化を実現する。local density の計算には単純に kd 木を利用し、dependent point の計算では kd 木を組み立てながら最近傍検索を行うというアイデアを取り入れることで、無駄なデータアクセスを削減した。また、いくつかの想定を置くことで o(n^2)時間でクラスタリングできることを示した。このアルゴリズムは早いものの、dependent point の計算にマルチスレッディングが利用できないという弱点がある。そのため、local density および dependent point の計算共にマルチスレッディングが利用可能な近似アルゴリズムも設計した。このアルゴリズムは厳密解と同じクラスタセンタを保証するため、高精度なクラスタリング結果を厳密解よりも高速に返す。またスレッド数に(ほぼ)比例して実行時間が短くなる。最後に、グリッドサンプリングを用いた近似アルゴリズムを設計し、ほぼ線形時間でクラスタリングできることを示した。実データを用いた実験から、既存のアルゴリズムに比べて大幅な性能向上を確認した。本研究の内容は SIGMOD2021 にて発表しており、5 で示す通り、本プロジェクトの主な成果の一つである。また、GitHub にてソースコードも公開している。

設計したアルゴリズムはユークリッド空間を想定して kd 木を使っているが、kd 木を別のデータ構造(vp 木や cover 木)に置換することで(計算量は O(n^2)となるものの)同様の操作が可能となり、効率的なクラスタリングがメトリック空間で実現できる。

(ii)「動的データに対する密度ベースクラスタリング問題」

この研究も Density-Peaks Clustering を考えており、データ集合が任意のデータの追加および削除によって動的に更新される環境を想定した。この環境において正確なクラスタリング結果をモニタリングする設定を考えたところ、メトリック空間では1つのデータの追加または削除において、Ω (n)時間かかることを証明した。つまり、厳密解を得るためには最低でも n(に比例する)個(数)のデータにアクセスする必要があり、データ集合が頻繁に更新される環境にスケールしない。また、Ω (n)時間よりも高速かつ精度保証を可能とする近似アルゴリズムも存在しないことを証明した。この理論的結果から、ヒューリスティックな近似アルゴリズムのみが実践的であることを示し、実践的に高速な近似アルゴリズムを設計した。このアルゴリズムは始点からの経路の長さがおおよそ O(log n)となる近接グラフを利用することで、おおよそ polylog(n)時間でクラスタリング結果を更新する。実データを用いた実験の結果から、提案アルゴリズムは既存の厳密アルゴリズムよりも 10 万倍以上高速であり、既存の近似アルゴリズムよりも高精度かつ 16 倍以上高速であることを確認した。本研究の内容は IEEE BigData 2022 にて発表しており、GitHub にてソースコードも公開している。

(iii)「距離に基づくアウトライア検出問題」

距離の閾値 r が与えられた時、あるデータ p から距離 r 以内に存在するデータの個数が k 未満であれば、p をアウトライアと定義する。この定義において全てのアウトライアを検出する問題に取り組んだ。この問題は、Apatch IoTDB にも実装されているものである。メトリック空間において高速にこの問題を解くために、これまでにいくつかのアルゴリズムが提案されているものの、既存アルゴリズムは全て O(n^2)時間かかってしまい、大規模データにスケールしない。この課題を解決するため、近接グラフを利用した新しいアルゴリズムを提案した。近接グラフは、各データが自身との類似度が高いデータと辺を作るデータ構造である。つまり、あるデータが

アウトライアかどうか評価する時、このデータから幅優先探索の要領で距離が r 以内に存在するデータを探索すれば、O(k)時間でインライアをフィルタリングできる。また、フィルタリングされなかったデータに対してのみ（線形スキャン等で）アウトライアかどうか厳密に計算すればよい。これにより、O((t+f)n)時間でアウトライアを検出できるようになる。（フィルタリングとアウトライアの厳密計算は単純な並列化が可能であり、マルチスレッディングにより計算時間を短縮できる。）

このとき、厳密な計算を必要とするデータの数（f）をできる限り小さくすれば、実行時間が削減される。つまり、インライアの検出率を最大にする近接グラフが望ましい。本研究では、任意の r と k に対して、距離 r 以内に存在する全てのデータにできるだけアクセスできる近接グラフを設計した。実データを用いた実験により、提案アプローチの性能は既存アルゴリズムのものよりも大幅に高いことを確認した。また、提案した近接グラフは既存の近接グラフよりも f を小さくできることを確認した。この内容は SIGMOD2021 にて発表している。

また、k 最近傍までの距離が最も遠い N 個のデータをアウトライアと定義する問題も存在し、いくつかのアルゴリズムが提案されている。しかし、これらも O(n^2)時間かかるという課題がある。この問題に対しても近接グラフを用いたアルゴリズムを提案し、提案アルゴリズムの時間計算量は O((N+f)n)であることを示した。また、実践的に N > f であり、実験的には O(Nn)時間となる。本問題は Ω（Nn）時間必要であることから、提案アルゴリズムの最適性を示している。実データを用いた実験により、提案アルゴリズムは既存アルゴリズムを超える性能を示した。この成果は The VLDB Journal にて発表しており、GitHub にてソースコードも公開している。本内容も、本プロジェクトの主な成果の一つである。

（iv）「高次元空間における近似 k 最近傍検索問題」

k 最近傍検索はコンピュータサイエンスにおける最も重要な操作の一つであり、多くの研究者が取り組んでいる。また、埋め込みベクトルの普及に連れて、高次元空間での最近傍検索の需要が高まっている。高次元空間では厳密解の計算が遅いことから、近似アルゴリズムの設計が盛んに行われており、近接グラフを用いたアルゴリズムが精度と速度のトレードオフが最も良いことが報告されている。近接グラフはある始点から探索を開始し、クエリに最も近いデータ（頂点）を貪欲に辿ることで近似 k 最近傍を検索する。しかし、始点の選び方によっては経路が長い場合や、k 最近傍データまでたどり着けないという問題がある。一方、始点と k 最近傍が近ければ高精度な解を高速に検索できる。この観測から、クエリに近い始点を計算し、そこからグラフ探索を行うアルゴリズムを設計した。クエリに近い始点の計算は locality-sensitive hashing（LSH）と呼ばれるハッシュ関数を利用する。これは、距離が近いデータには似たハッシュ値を与える関数である。このハッシュ関数とサンプリングを利用することで、始点の計算を定数時間で行える。また、複数のハッシュ関数を利用することで、クエリに近い始点を得られる確率が指数関数的に増加することを示した。また、マルチスレッディングを利用することで複数のハッシュ関数を利用した場合でも1つのハッシュ関数を利用した場合と同じ時間で検索できる。

実データを用いた実験から、既存の近接グラフを用いたアルゴリズムよりも高速・高精度に k 最近傍検索を行えることを確認した。実験ではユークリッド距離を使っているが、提案アルゴリズムは LSH のサポートがある任意の距離関数が利用可能である。本研究の成果は DEXA2021 で発表しており、Best Paper Award を受賞している。また、本研究は Yahoo! Japan

研究所との共同研究である。GitHub にてソースコードも公開している。

(v)「k 最近傍距離推定問題」

局所密度の計算や(iii)の問題のように、k 最近傍までの「距離」のみ知れば十分な(k 最近傍データ自体は必要がない)応用が数多く存在する。この距離を計算する最も単純な方法は k 最近傍検索を行うことであるが、一般的に距離計算やデータアクセスは検索のボトルネックであり、距離のみを把握する場合にはコストが大きい。この課題を解決するため、機械学習を利用したアプローチを提案した。機械学習を用いて高精度な距離を推論するためには、一般的に複雑なモデルが必要である。しかし、複雑なモデルは計算コストが大きくなるため、そもそもの課題を解決できない。本研究で提案した方法はこのジレンマを解消した。

データ空間をグリッドで分割し、各セルの中心にピボット点を設ける。ピボット点の k 最近傍は事前に計算できることから、クエリとピボットの距離、ピボットの k 最近傍距離、および三角不等式を利用することで、セルが十分小さければ高精度な k 最近傍距離を推論できる。しかし、細かいセルは大量のメモリを利用するため現実的ではない。そこで、ニューラルネットワーク(MLP)を用いてセルが細かすぎない場合でも高精度な距離を推論できるように学習する枠組みを設計した。提案したモデルは k の最大値 k_{max}までの距離推定を定数時間で行い、非常に高速に距離推定を実行できる。

実データを用いた実験の結果、厳密解アルゴリズムよりも高速であり、既存の学習モデルを利用したものよりも高精度であることを確認した。また、4 つのケーススタディを行い、各アプリケーションに対して高精度な結果を高速に提供できることを示した。本研究も Yahoo! Japan 研究所との共同研究であり、SIGSPATIAL2022 で発表している。GitHub にてソースコードも公開している。

(vi)「逆最大内積探索問題」

ビデオオンデマンドサービス・EC サイト等はいつでもどこでも利用できるようになり、レーティングやクリック等のフィードバックを考慮することで、ユーザに高精度な推薦サービスを提供する推薦システムが普及した。推薦システムでは一般的にユーザやアイテムは高次元ベクトルで表現され、ユーザベクトルとアイテムベクトルの内積によってユーザとアイテムのマッチング度合いを推定する。例えば、あるアイテムの広告を行いたい場合や市場規模を調査したい場合、ど(れだけ)のユーザがそのアイテムを気にいるかを把握することが重要である。しかし、これまでの研究ではこの要求に応えられる問題が存在していなかった。そこで本研究では、クエリアイテムがあるユーザにとって最大内積探索の結果になるか(なる場合はそのアイテムはユーザにとって好ましいアイテム)に着目し、クエリが最大内積探索結果に含まれる全てのユーザを探索する、逆最大内積探索を定義した。

この問題を解く単純な方法は、全てのユーザに対して最大内積探索を行うものである。一般的に、ユーザとアイテムの数は大規模であるため、この方法では計算コストが大きい。この課題を解決するため、あるユーザが検索結果に含まれるか含まれないかを定数時間で計算する技術を開発した。これに加えて、検索結果に含まれない複数のユーザを一度に計算(フィルタリング)する技術も開発した。これらの技術により、単純に最大内積探索を繰り返す方法よりも数

百倍以上高速に検索結果を返すことに成功した。さらに、マルチスレッディングを利用することでさらに計算時間を削減することもできる。本内容は RecSys2021 および ACM Transactions on the WEB で発表しており、GitHub にてソースコードも公開している。本内容も、本プロジェクトの主な成果の一つであり、日本で毎年開催されている RecSys 論文読み会にも招待されて成果を紹介した。

(vii)「多様性を考慮した最大内積探索問題」

上で述べたように、最大内積探索はユーザの好みのアイテムを検索できるが、推薦リストに偏りが生じることも報告されている。WWW や KDD の基調講演でいくつかの企業から報告されているが、推薦リストの多様性はユーザの長期間のサービス利用に効果がある。これらの事実に基づき、多様性を考慮した最大内積探索問題を定義した。これは、ユーザベクトルとの内積を最大としつつ、検索結果内のアイテムの類似度を最小にするものである。また、多様性の度合いはパラメータとしてユーザが指定できる。ケーススタディを行い、例えばあるシリーズの映画しか推薦リストに入らないユーザの場合、多様性を考慮することで内積が高いまま様々なジャンルの映画が推薦リストに含まれるようになることを確認した。

本問題は NP 困難であるため、厳密解を多項式時間で得ることはできない。そこで、貪欲法を利用した近似アルゴリズムを設計した。貪欲法は、暫定解に対して目的関数を最大化するアイテムを暫定解に追加する操作を k 回繰り返す方法である。単純に貪欲法を用いると、O(nk) 回のデータアクセスが生じ、n(アイテム数)が大きい場合に低速となってしまう。この問題に対して、貪欲法と同じ解を保証しつつ、不要なアイテムベクトルへのアクセスをスキップする技術を開発した。この技術によりデータベースアクセスを early stop でき、データ数が大きい場合にもインタラクティブな推薦リストの作成が可能となる。

実データを用いた実験の結果から、単純な最大内積探索よりも多様性の度合いが高いことを確認した。また、提案アルゴリズムは単純な貪欲法よりも 10 倍以上高速に解を計算できることを確認した。本内容は RecSys2022 で発表しており、GitHub にてソースコードも公開している。また、RecSys 論文読み会にも招待されて成果を紹介した。本研究も Yahoo! Japan 研究所との共同研究である。

3. 今後の展開

本プロジェクトで取り組んだ問題は、これまでに取り組まれていた、既存のシステムに実装されていた、または要求はあったものの未実装なものであり、既に多くの応用がある。本プロジェクトで提案したアルゴリズムはソースコードを公開していることもあり、既存の実装との置換は容易であると考える。そのため、何らかのサービスが本プロジェクトで開発した技術を使いたい場合は任意のタイミングで可能であり、社会実装の準備は整っている。

近年、コンピュータサイエンス分野において公平性に注目が集まっている。これは、本プロジェクトで取り組んだ問題を背景とすると、結果に含まれるデータが特定のグループに偏っていないか、各データが結果に含まれる確率に偏りがないか、といった概念である。例えば、各データが複数ある中のある1つのグループに属しているとする。この想定で距離に基づくアウトライア検出問題を考えた時、アウトライアとなるデータはマイノリティとなるグループに属しているデータとなる可能性がでてくる。直感的にはこれらのデータはアウトライアというわけでは

ないため、グループの分布とグループ間の公平性を考慮する必要が出てくる。このように、社会が結果の公平性をデフォルトの要件とするような時代となった場合、本プロジェクトの成果をそのまま社会に実装できるわけではなくなる。公平性は全世界共通の認識であるため、本プロジェクトで取り組んだ問題に公平性という制約を加えたものに取り組み、同様な結果を数年で出す必要性があると考えている。

4. 自己評価

多くの応用があるプリミティブオペレータに対するメトリック空間で高速かつ並列化可能なアルゴリズムを設計する、という本プロジェクトの主な目的は達成できた。また、主要な成果は全てオープンソース化し、GitHubからMIT License下で誰でも利用可能なように整備したことからも、アルゴリズム基盤の開発はできたと考えている。

プロジェクト期間の途中からコロナ禍となり、予定していた出張は全て無くなったが、その分実験や開発環境に研究予算を割り当てることができ、研究の進捗という意味ではコロナの影響を受けなかったように思う。また、主要な論文についてはオープンアクセスにしており、多くの研究者の目に触れるように心がけた。

本プロジェクトで取り組んだ問題は比較的クラシックなもので、多くの研究者が取り組んでいる。また、論文やコードをオープンにすることで後続の研究に発展しやすい点、およびエンジニアにも利用可能である点から、本プロジェクトの成果は今後の研究のベースラインとして扱われる効果を有する。また、提案したアルゴリズムは 1 台の計算機で高速に動作し、マルチスレッディングでさらに高速化されることから、複数の計算機で構成される分散システムに対して経済的、管理的、および環境的に優位な位置にあるため、計算機システムに投資するコストを削減し、CO2 排出量を削減するメリットを有する。

最後に、本プロジェクトの成果の多くは、トップ国際会議・国際論文誌（SIGMOD、RecSys、SIGSPATIAL、AAAI、The VLDB Journal、ACM Transactions on the WEB）で発表できた。つまり、これらの成果は国際的に高い評価を得られたことを客観的に示している。また、データベース、推薦システム、および位置情報システムといった様々な分野で評価されていることも分かる。これらの事実からも日本の技術レベルのプレゼンスを示すことに貢献できたと言える。

5. 主な研究成果リスト

(1) 代表的な論文（原著論文）発表

研究期間累積件数：34件

1. Daichi Amagata and Takahiro Hara, Fast Density-Peaks Clustering: Multicore-based Parallelization Approach, In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 49-61, 2021.	

本論文では、多次元ユークリッド空間における密度ベースクラスタリング問題に取り組んだ。あデータの数をnとすると、既存研究で提案されているアルゴリズムではO(n^2)時間必要であった。本論文で提案したアルゴリズムは、一定の条件を満たす際にo（n^2）時間で計算可能であることを示した。実データを用いた実験から、提案アルゴリズムは既存アルゴリズムよりも大幅に性能が改善されていることを確認した。	
2. Daichi Amagata, Makoto Onizuka, and Takahiro Hara, Fast, exact, and parallel-friendly outlier detection algorithms with proximity graph in metric spaces, The VLDB Journal, volume 31, number 4, pages 797-821, 2022.	
本論文では、メトリック空間における距離に基づくアウトライア検出問題に取り組んだ。既存アルゴリズムは O(n^2)時間必要であるが、本論文では近接グラフを用いた新たなアプローチを提案し、提案アルゴリズムは O((t+f)n)時間で動く（t はアウトライアの数、f はフィルタリングされなかったインライアの数）。実データを用いた実験から、提案アルゴリズムは既存アルゴリズムよりも大幅に性能が改善されていることを確認した。	
3. Daichi Amagata and Takahiro Hara, Reverse Maximum Inner Product Search: How to efficiently find users who would like to buy my item?, In Proceedings of the ACM Recommender Systems Conference, pages 273-281, 2021	
本論文では、逆最大内積探索という新たな問題を提案した。これは、あるアイテムを好みそうなユーザの集合を探索するもので、広告配信や市場調査等の応用がある。この問題は最大内積探索問題をすべてのユーザに対して解くことで解を計算できるが、非常に遅い。この課題に対して、解に含まれないユーザを定数時間でフィルタリングできる技術を開発した。実データを用いた実験から、提案アルゴリズムはベースラインよりも 500 倍以上高速であることを確認した。	

（2）特許出願

該当なし。

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

- Best Paper Award, International Conference on Database and Expert Systems Applications (DEXA) 2021
- 企業（LegalForce）賞、第 14 回データ工学と情報マネジメントに関するフォーラム（DEIM フォーラム 2022）
- 2022 年度マイクロソフト情報学研究賞、情報処理学会