研　究　報　告　書

「大規模データに対する高速類似性解析手法の構築」

研究タイプ：通常型
研究期間：　平成 21 年 10 月～平成 25 年 3 月
研 究 者：　宇野 毅明

1.　研究のねらい

巨大データを解析するには様々な問題がある。データが初等的でモデル化が難しく、データを解析しようと思っても、何ができるか、何をすればよいのか分からないことが多い。そのまま可視化をする、そのまま知識発見を行うのは一般に難しい。一方で、近年盛んに解析されているネットワークなどのデータでは、物と物との関係性に着目した解析が行われている。このような解析では、データのサンプリングを行ってしまうと、関係が失われてしまうことが多く、精度の高い解析が不可能となる。また、サンプリングをすると、比較的小さなまとまりはノイズと区別できなくなってしまう、ということもある。このような点から、巨大データをサンプリングして小さくしてから解析するという方法は有効でない。巨大データの関係性を直接調べ、それを元にした解析手法を実現することが肝要である。

本研究では、関係性の中でも比較的取り扱いが容易な「類似性」に着目する。一つ目の研究課題は、大規模データの類似性を網羅的に調べる効率良いアルゴリズムを構築することである。データが大規模であると、全ての項目の組について類似度を調べるのは非常に大きなコストがかかる。そこで、類似する物のみに対して、類似度を調べるという問題設定を行い、その元で高速なアルゴリズムを開発する。これにより、ある程度一般的なデータに対して、基礎的な類似性を高速に計算できるようになることを目標とする。大規模データの解析を行いたいときには、このような基礎的な計算では不十分なことが多く、このような計算結果らより具体的、あるいは意味的な構造を見つけ出して可視化する必要がある。そこで、本研究では、そのような有効な解析手法についても研究し、幾つかの具体的なターゲットとなるデータとタスクを選び、それらに対して、類似性の計算結果を用いた解析手法の開発を行う。

2.　研究成果

（1）概要

巨大データ解析のため、大規模データの類似性を網羅的に調べる効率良いアルゴリズムを構築することをテーマとした。大規模データの解析を行いたいときには、計算結果らより具体的、あるいは意味的な構造を見つけ出して可視化する必要がある。そこで、可視化の有効な解析手法についても研究し、幾つかの具体的なターゲットとなるデータとタスクを選び、それらに対して、類似性の計算結果を用いた解析手法の開発を行った。

本研究では、まず「文字列パターンが現れる場所の組合せ」を特定することとし、非常に短い時間で、今までは考えられなかったような長いパターンをたくさん見つけるアルゴリズムを開発することに成功した。これにより長い頻出文字列パターンのマイニングが初めて可能となった。

文字列の類似性を解析するアルゴリズムを文字列の圧縮アルゴリズムにも適用し、最適な

選択を行うアルゴリズムの開発に成功した。

グラフの列挙アルゴリズムについても研究を行った。類似するグラフの同型性の判定を大幅に削減する手法を考案し、実装を進めている。現在はより高速なプログラムを実装すべく開発を継続している段階である。

他の類似尺度については、まだ決定的な成果を得られていない。今後、他の類似性尺度、および包含や対立など他の関係性について網羅的な計算を高速で行う手法の研究を進めていく。また、実際のデータ解析についても研究を展開していく。

（2）詳細

類似性の解析を行う基礎的な手法については、さきがけ研究前にプロトタイプの開発が終了していた文字列の類似性を解析するアルゴリズムを一つの基盤として開発を行った。まず、一つ目の課題として、文字列のアルゴリズムを利用して、他の構造を持つデータの類似性を調べる高速アルゴリズムの開発を行った。

類似性を解析する方法として、Local Sensitive Hash(LSH)という方法が知られている。これはある種のランダムに発生した関数を用いて、データを文字列符号に変換するというもので、類似する2つの項目の符号が高い確率で（ハミング距離の意味で）短い距離を持つようになる。通常、LSH は「同一の LSH 符号を持つ項目の対」を比較することで、類似しない項目対の比較を省略する。しかし、類似する2つの項目がまったく同一の LSH 符号を持つことはやはり確率が低く、その確率を上げるために異なる LSH 符号を非常に沢山作る、ということが行われており、これが LSH を使った手法の効率を著しく下げていた。今回の研究では、「同一の LSH 符号を持つ項目の対」だけでなく、「類似する LSH 符号を持つ項目の対」を比較することで LSH 符号の精度を高め、効率性を上昇させることに成功した。従来は類似する LSH 符号を持つ項目の対を見つけることが高いコストを要求したが、以前開発した文字列の類似性を解析するアルゴリズムを用いることで、短い時間で終了することを可能とした。これにより、同じ精度で計算した場合、場合によってはメモリ効率が 10 倍、速度が 100 倍になる程度の改良を実現した。

文字列の類似性は、ゲノム配列を初めとする、中規模な類似構造を持つような文字列の距離計算にも適用した。文字列の類似性を計算する場合、編集距離と呼ばれる、2つの文字の対応を取るために置換、挿入、削除の回数をどれだけ少なくできるか、という尺度が類似度として用いられることが多い。しかしこの距離は計算に時間がかかり（長さの 2 乗）大規模な文字列データでは適用が難しい。今回は、中規模な類似構造、つまり全体の 1000 文字から全体の 10%程度の長さの類似構造をいくつか持つような文字列の類似性を調べる場合に焦点を当て、新しいアルゴリズムの開発を行った。このような類似構造を持つ場合、その類似構造は距離計算を行う際に「対応する」と見なされるべきである。ただし、複数の場所と類似する部分は、その中のどれか1つとのみ対応することになる。このような観察から、まず最初に中規模な類似構造を計算し、全体的な類似度を計算する際には、どの中規模構造をつなげるべきか、という点にのみ焦点を絞って距離計算をする、というモデルを考案した。中規模な類似構造の発見には、先の文字列の類似性を解析するアルゴリズムを用いる。これは非常に短い部分列の類似性のみを発見するので、見つかった短い類似構造を進展することで中規模の類

似構造を得ることとした。全体の計算に関しては、新しいアルゴリズムを構築し、O(knlogn)、あるいは O(n√n)(n は中規模構造の数、k は類似度の中で局所的に許す異なりの数)の時間で動くようなアルゴリズムを開発した。実験的にもパフォーマンスは良好で、大規模な文字列データでも実時間で比較をすることができた。

同じような発想で、データマイニング界の難問とされる、頻出文字列マイニング問題に対しても取り組みを行った。文字列マイニングは、文字列データの中に多数現れる文字列パターンを見つける問題であり、この際、多少エラーを許す、つまり似た文字列がデータ沢山現れるような文字列パターンを全て見つける問題となる。山登り探索といういわゆる既存手法に基づいて問題設定を行うと、短めのパターンがほぼ全て頻出パターンとなり、アルゴリズムはそれらを全て探索する必要があるために非常に遅くなる。そのため、既存手法を用いることができず、長らく未解決問題であった。

今回の研究では、文字列パターンが現れる「場所の組合せ」に注目した。パターンが現れる際に類似性を許していることから、異なるパターンがまったく同じ場所(の組合せ)に現れることがありうる。このような場合、本来パターンは一つ出力すれば十分なはずである。この観察に基づき、本研究では、まず「文字列パターンが現れる場所の組合せ」を特定することとした。具体的には、文字列データから類似する場所を全て見つけ出し、たくさんの場所と似ている場所を抽出することとし、これら「似ている場所」が作り出すグループからパターンを生成することとした。結果、非常に短い時間で、今までは考えられなかったような長いパターンをたくさん見つけるアルゴリズムを開発することに成功した。実際に見つかったパターンはデータベースに多く現れており、これにより長い頻出文字列パターンのマイニングが初めて可能となった。

文字列の類似性は、文字列の圧縮にも用いてみた。ゲノムなどのランダム性を含有する文字列データは、そのランダム性ゆえに通常のハフマン符号や繰り返し構造に基づく方法ではほとんど圧縮できない。その一方で、ゲノム配列などは、比較的大きな類似構造を持つことも分かっている。類似構造があるのであれば、似ている物が存在する部分については、「〇〇の部分にこういう変更を加えた」という形で表現することで、比較的コンパクトに圧縮することができる。文字列の類似性を解析するアルゴリズムを用いれば、このような類似性は網羅的に見つけることができるので、これを使って圧縮アルゴリズムを作ることとした。1つの部分は複数の部分と類似することがあり、また類似構造の始まり、終わりもまちまちである。その中からどの類似構造を使うと効率が良くなるか計算するため、動的計画法という手法を用いたアルゴリズムを開発し、最適な選択を行うアルゴリズムの開発に成功した。

現在最も良いとされている圧縮アルゴリズム、および過去に提案されてきたゲノム圧縮アルゴリズムは、圧縮によって 10%程度の削減しかできなかったところを、今回のアルゴリズムは 20%近い削減を可能とした。

グラフの列挙アルゴリズムについても研究を行った。グラフでは、等しいグラフは同型であると呼ばれ、同一視されるが、同型性の判定は難しい問題として知られている。そのため網羅的に列挙することもまた難しかった。似たグラフが同型であるかどうかを判定し、重複を回避するのにコストがかかるのである。今回は、類似するグラフの同型性の判定を大幅に削減する手法を考案し、それをもとに実装を進めている。予備実験により大幅な計算の省略ができることが確認できている。現在はより高速なプログラムを実装すべく開発を継続している段階で

ある。

3.　今後の展開

今回の研究で、文字列を中心とした類似性の解析、および類似性を用いたデータ解析手法については、いくつかのマイルストーンができたと考えている。しかしながら、他の類似尺度については、まだ決定的な成果を得られていない面は否めない。今後は他の類似性尺度、および包含や対立など他の関係性について網羅的な計算を高速で行う手法の研究を進めていく必要がある。また、解析手法がある程度整いつつあることを鑑み、実際のデータ解析についても研究を展開する必要がある。近年ビッグデータと呼ばれるビジネス・研究分野が盛んであるが、このような解析に高速アルゴリズムを必要とする分野では、アルゴリズムを基礎分類の軸としたような系統立てを持つような研究の俯瞰が必要であると考えている。ビッグデータの解析に関して、どのようなアルゴリズムがどのような課題、データに対して有効であるのかを解明し、それを元にしてモデリングを行う標準的、あるいは発展的なモデリング手法を系統立てる。また、それに伴い、各種のビッグデータ自身の性質に関しても、計算の側面、つまりどのようなアルゴリズムが有効でどのような手法がどうような精度を持つのか、という側面に関して明らかにしていく必要があると考えている。データの性質と計算手法の性質、両面からの研究を進めることで、より迅速かつ明解に、ビッグデータ解析の研究が進んで行くであろう。

4.　自己評価

研究開始時、当初の目標はアルゴリズムの発展であった。文字列の類似性を解析するアルゴリズムのプロトタイプがあったため、このアルゴリズムの完成度を高めること、他の類似性・関連性の計算へ拡張すること、あるいは他のアルゴリズムを考案すること、類似性の解析を用いたアプリケーション、つまり計算結果を使った新しいデータ解析手法を開発することであった。アルゴリズムの研究では、アルゴリズムの開発は単発で終わることが多く、実際に実装を作成して応用分野で使うこと、およびそのアルゴリズムを用いたアルゴリズムを開発する、という形の縦横の展開を持つものは少なかった。その意味で今回の研究は当初目的を達成し、意義ある物となったと考えている。

研究が終了して鑑みると、このようなアルゴリズムの展開の他に重要な課題が見えてきた。データ解析においては、単純・基礎的・効率の良いアルゴリズムとそれを用いた解析手法の存在が重要であると共に、それらアルゴリズムを用いたモデルの開発技術や、計算やアルゴリズムの側面から見たデータの性質（どのような手法でどのような精度がでるか）といった点も重要であることが分かってきた。現在ビッグデータの解析においては、様々な手法やデータが乱立しており、さながら戦国時代のようである。このような時分において、データ解析手法の研究に対して明解な方向性を得ることができる概念を得られたことは、今後の研究の展開に対して大きな成果であると考えている。

5.　研究総括の見解

大規模データ内の汎用な事例間類似性高速計算アルゴリズムの開発が研究課題である。

今回の研究で、文字列を中心として、非常に短い時間で、今までは考えられなかったような長いパターンをたくさん見つけるアルゴリズムを開発することに成功し、長い頻出文字列パターン

のマイニングが初めて可能となっている。また、これを文字列の圧縮アルゴリズムにも適用し、最適な選択を行うアルゴリズムの開発に成功している。

今後、他の類似性尺度などの関係性についても手法の研究の範囲を拡大していってほしい。また、実際のデータ解析についても研究を展開していってほしい。

6. 主な研究成果リスト

(1)論文(原著論文)発表

1. 著者、発表論文タイトル、掲載誌名、巻号頁、発行年等
Takeaki Uno, Ryuhei Uehara, Shin-Ichi Nakano: Bounding the Number of Reduced Trees, Cographs, and Series-Parallel Graphs by Compression, WALCOM 2012, Lecture Notes in Computer Science 7157, pp. 5-16 (2012)
2. Yasuo Tabei, Takeaki Uno, Masashi Sugiyama, Koji Tsuda: Single versus Multiple Sorting in All Pairs Similarity Search, Journal of Machine Learning Research - Proceedings Track 13, pp. 145-160 (2010)
3. Takeaki Uno: Multi-sorting algorithm for finding pairs of similar short substrings from large-scale string data. Knowledge Information Systtems 25(2): 229-251 (2010)
4. 松井鉄史、宇野毅明, 計算幾何学的な手法を用いた高速相同性計算手法, 情報処理学会バイオ情報学研究会 (2010)
5. Takehiro Ito, Shin-Ichi Nakano, Yoshio Okamoto, Yota Otachi, Ryuhei Uehara, Takeaki Uno, Yushi Uno: A Polynomial-Time Approximation Scheme for the Geometric Unique Coverage Problem on Unit Squares. SWAT 2012, Lecture Notes in Computer Science 7357, pp. 24-35 (2012)

(2)特許出願

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)

受賞:文部科学大臣表彰 科学技術部門 若手科学者賞受賞.

巨大データ解析に対する超高速アルゴリズム構築法の研究 (2010 年 4 月 13 日)

招待講演: Deep of Enumeration Algorithms, ENUMEX(ヨーロッパの研究者向けの集中講義), イタリア (2012 年 9 月 26 日)