距離木とk近傍グラフを用いた超高次元データの近傍検索

YUSUF MUKARRAMAH　渡辺知恵美　瀬々潤
お茶の水女子大学大学院人間文化創成科学研究科理学専攻

1　はじめに

近年，遺伝子発現量の検索エンジンの要求が高まっている．医療では，遺伝子発現量の検索エンジンは，医師の診断を助けることができる．例えば，医師が患者から採取した遺伝子発現量を入力し，過去のどの患者の遺伝子情報に近いかを検索することで，薬に対する副作用の可能性を調べることができる．

文献 [5] では，これらの問題を考慮した遺伝子発現量の検索エンジンのプロトタイプがwebアプリケーションとして開発されたが，素早い検索機能が求められている．そこで，本研究は，このアプリケーションを高速化するための手法を考える．

我々は，M-Tree[1] と k 近傍グラフ [6] を用いた手法を提案した．それに対して，人工データと実データを用いて実験を行った．実験で提案手法の有用性を確認することができた．更に，リアルタイム検索ができるように，四分木を使った手法 [8] も提案した．

2　超高次元データを扱う時の問題

遺伝子発現量データの索引を構築するときは，2つの問題を考えなければいけない．一つ目は，遺伝子発現量データは超高次元データである．超高次元データを扱うときに生じる問題として，類似性を一般的なユークリッド距離で表すと，次元の呪いに陥ってしまうということである．

2つ目の問題は，多様体型データを考える必要がある．データの次元は５０以上になると，データ分布が多様体を構成すると知られている．この多様体上での検索の方が有意義である．ユークリッド距離的に近いとみなしても，多様体上では遠かったら類似関係が弱いと考える．

1つ目の問題に対して，M-Tree[1] を利用することを考えた．M-Tree[1] は膨大な超高次元データを管理・アクセスするための索引である．ユークリッド空間やベクトル空間上では表現できないもののために提案された．いわゆる距離空間 (非負性，対称性，三角不等式を満たす空間) の索引である．

しかし，M-Treeのみで検索を行う場合，ロールケーキデータ [4] のような，グラフ上での距離の方が有意義なデータに対して，遺伝子発現量検索が失敗してしまう．そこで，k 近傍グラフ [6] を用いた手法を提案した．

k 近傍グラフは多様体状のデータのクラスタリングに用いられる．その研究の一つは，Isomap アルゴリズム [4] である．Isomap は，近傍グラフを用いてデータ間の類似関係を表した後、データをユークリッド空間へマッピングし、クラスタリングを行った．このアルゴリズムは，ロールケーキ状のような特別な多様体にも成功した．

しかし、遺伝子発現量検索エンジンの場合，元の多様体に新しい点が加わることとなるため，データをIsomapでマッピングしても，マッピング結果が変わってしまう．そのため、我々は、データをマッピングせずに，k近傍グラフのみを利用することにした．

3　超高次元データに対する高速な検索のための手法と評価実験

超高次元データに対する検索を２つのケースに分けて考えた．超高次元データのみを考慮する場合，確立された技術であるM-Treeを利用することができる．そして，多様体型であることも考慮する場合，k 近傍グラフとM-Treeを併用した検索手法を提案する．

3.1　M-Tree と k 近傍グラフを合わせた方法

図１は提案手法の処理の流れを示す．まず，前処理として，超高次元の点データのM-tree と k 近傍グラフを作成し，保存しておく．次に，ユーザがクエリ点 q の l 近傍を問い合わせる時，q を k 近傍グラフに組み込む（図１（1），(2)）．そして，q が組み込まれた k 近傍グラフ上で検索を行う．k 近傍グラフ上で，l 近傍を検索をするためには，ダイクストラ法を用いる．ダイクストラ法は、本来最短経路を求めるためのアルゴリズムだが，アルゴリズムからして，出発点から近い順に点を走査していくので、結果的に l 近傍（図１（3））が求められるということになる．

q を k 近傍グラフに組み込む操作には，二つの処理をする必要がある．(1)q の k 近傍を求め，そして q からそれらの点へ辺を張る．(2)k 近傍グラフのあるノードが q がその点の k 近傍に含まれたら，その点の k 番目の辺を削除し q へ辺を辺の繋げ換えを行う．

ダイクストラ法でグラフをスキャンしながら処理 (2) を実行する手法として，２つの手法を考えた．手法１は，ノードをスキャンする際 q が近傍かどうかをチェックし，そうである場合辺の繋げ換えを行う．q への距離計算を

[0]Finding Nearest Neighbors for A High Dimentioanal Data by Using Metric Tree and knn Graph
[1]Mukarramah Yusuf, Chiemi Watanabe, and Jun Sese, Graduate School of Humanities and Sciences, Ochanomizu University

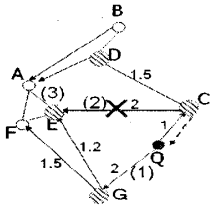

図 1: k 近傍グラフ上で l 近傍を求める（k=2, l=4）

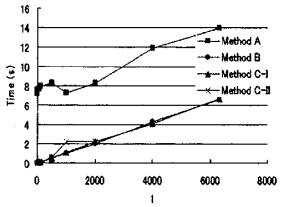

図 2: 人工データの実験結果 (k=10)

最小限に減らす方法として，手法 2 を提案する．手法 2 は，k 番目の辺を最短パスとして通ろうとする時のみ，処理 (2) のチェックを行うので，計算コストが削減できる．

3.2 評価実験

提案手法の有用性を測るために，人工データと遺伝子発現量データを使用し，評価実験を行う．実験では，次のそれぞれのグラフ上で検索を行った結果を比較・検討する：(1)Method A：クエリ点を含んだグラフ，(2)Method B：クエリ点から差分的に辺を張ったグラフ，(3)Method CI：クエリ点から差分的に辺をつなぎ換えたグラフ（手法 1 で），(4)Method CII：クエリ点から差分的に辺をつなぎ換えたグラフ（手法 2 で）．

人工データは生成したロールケーキ状の多様体を構成する 6308 個の点である．それぞれのデータは x 軸，y 軸，z 軸にあたる 3 つの属性を持っている．実データは 447 次元を持つ遺伝子発現量データである．パラメータ k を 10 に設定し，それに対して l を 1 から 6308 までの値に変化させた．

各 l に対して，10 個の異なる点データで検索を行い，再現率，適合率，時間の平均を求めた．Method C-I と C-II に対して，更に，チェック回数も記録した．再現率と適合率はそれぞれ，Method A と比較したものである．実験結果の一部は図 2，3 に示す．

4 リアルタイム検索への試み

提案手法の計算において，データの k 近傍グラフをディスク上に保存しておき，クエリがあった際にメモリに読み込み，l 近傍を求める．実際の応用では，扱うデータがサイズがとても大きく，それによって，読み込まなければいけない k 近傍グラフも大きくなり，メモリに乗り切れないことが予測される．

更に，提案した手法において，計算をする際，チェックと呼ばれる処理を行わなければいけない．データの数が増えると，チェック回数も増え，時間コストがかかってしまい，リアルタイム検索ができなくなる可能性がある．

よりリアルタイムな検索を実現するため，我々は四分木を用いた手法を提案した [8]．この手法は Hanan Samet らが提案した SILC と名づけられた手法 [7] の拡張である．

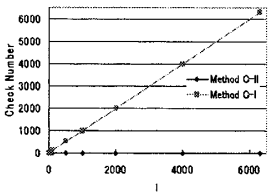

図 3: 遺伝子発現量データの実験結果 (k=10)

5 おわりに

本研究の目的は超高次元データの検索を高速化することである．超高次元データに対する検索を 2 つのケースに分けて考えた．超高次元データのみを考慮する場合，M-Tree を利用することができる．そして超高次元データが多様体をみなす場合，k 近傍グラフを利用する手法を提案した．提案手法に対して，検証実験を行った．超高次元データの類似検索に k 近傍グラフは利用できるという結果が得られた．

そして、リアルタイムな検索のため，四分木を用いた手法を提案した．この手法は，MethodB のように，辺の削除をしないことを前提として考えたため，辺の削除が必要となった時の高速化の方法を考える必要がある．

参考文献

[1] P. Ciaccia, M. Patella and P. Zezula: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces, VLDB '97, 426-435, 1997.

[2] R. Harpaz and R. Haralick: Exploiting the Geometry of Gene Expression Patterns for Unsupervised Learning, ICPR '06 Vol 2, 670-674, 2006.

[3] R. Souvenir and R. Pless: Manifold clustering, ICCV '05, 648-653, 2005.

[4] J. B. Tenenbaum, J. C. Langford and V. de Silva: A Global Geometric Framework for Nonlinear Dimensionality Reduction, SCIENCE Vol 290 No 9, 2319-2323, 2000.

[5] 梅澤香矢乃, 瀬々潤: MARE: 遺伝子発現量検索エンジンの構築に関する一考察, DEIM '00, 2009.

[6] T. B. Sebastian and B. B. Kimia: Metric-based shape retrieval in large databases, ICPR '02, 291-296, 2002.

[7] H. Samet, J. Sankaranarayanan and H. Alborzi: Scalable Network Distance Browsing in Spatial Databases, SIGMOD '08, 42-54, 2008.

[8] ユスフムカルラマー, 渡辺知恵美, 瀬々潤: 高次元データのリアルタイム検索への試み, DEIM '10, 2010（発表予定）．