高次元データ空間における大規模近傍検索のための近似 k 最近傍グラフに関する研究

2013 年度

岩崎 雅二郎

主　論　文　要　旨

報告番号	甲 ㊙ 第　　号	氏　名	岩崎 雅二郎

主 論 文 題 目 :

高次元データ空間における大規模近傍検索のための
近似 k 最近傍グラフに関する研究

（内容の要旨）

スマートフォンの普及に伴いインターネット上にはマルチメディアデータを含む多様な情報が急激に蓄積されるようになり，情報検索技術への要求が急増している．マルチメディアデータを検索する場合にはデータの内容を表す特徴量を抽出し，元データの代わりに，その特徴量のデータを検索する．検索精度を高めるためには高次元データが必要となるので，高次元データを高速に検索することが可能な距離基準空間インデックスが不可欠である．k 最近傍グラフをインデックスとして用いた手法は他の従来手法よりも検索時の計算量を削減できる．k 最近傍グラフは各ノードがエッジによって k 個の近傍ノードへ接続されており，任意のノードから指定された検索点に向かって，より近いノードを順次たどることで，近傍ノードを検索することができる．しかし，この検索手法には 2 つの課題がある．第 1 に，少ない計算量で検索が可能である反面，k 最近傍グラフの生成時に膨大な計算量が必要となる．第 2 に，グラフが分離する場合には候補ノードに到達することができず検索精度（再現率）が抑制される．

本論文では木構造型インデックスを組み合わせた近似 k 最近傍グラフインデックスを提案する．k 最近傍グラフの生成時には各ノードの近傍ノードを特定しなければならず，この処理に多大な計算量が必要となる．提案手法ではノードをグラフに逐次追加し，追加対象のグラフを用いて追加ノードの近傍ノードを検索することで，グラフ生成時の計算量を大幅に減らすことができる．また，逐次ノードをグラフに接続することによりグラフが分離することがなく検索精度の抑制が解消される．さらに，検索時のグラフの探索では任意のノードではなく，木構造型インデックスにより近傍ノードを特定した上で，そのノードからグラフの探索を開始することにより，検索時の計算量の削減が可能となる．一方，提案手法ではエッジが過剰に付与されるノードが生成される傾向があるので，エッジを削減する必要がある．この際，グラフの連結性を維持するために，エッジの削除によりグラフが分離するか否かを判定した上で分離しない場合には過剰なエッジを削除し，分離する場合にはエッジを削除せずに過剰ノードから遠方のノードへ削除対象エッジを移動する手法を提案する．

10 万オブジェクトの 50 次元一様分布データにおいて k 最近傍グラフに対して提案手法では生成時に 96.7%の計算量が削減された．10 万オブジェクトの画像特徴量（1,228 次元）およびエッジ数 k が 16 において検索精度 95%を得る場合に k 最近傍グラフに対して提案手法では約 73%の計算量が削減された．さらに，百万オブジェクトの画像特徴量に対してエッジ削減の手法によりエッジ数が最大 34.2%削減された．以上より，提案手法によって計算量とエッジ数を削減できることを確認した．最後に，提案手法を実際の商品画像検索に適用する事例を紹介する．

SUMMARY OF Ph.D. DISSERTATION

School	Student Identification Number	SURNAME, First name IWASAKI, Masajiro

Title

A Study on Approximate *k*-Nearest Neighbor Graph for Large-scale Proximity Search in High-dimensional Data Space

Abstract

As the smartphone has become common recently, huge multimedia data has been accumulated on the Internet, causing the demands of information search technology to grow rapidly. To search for the multimedia data, the features representing their contents are extracted from the data, and then the features are searched for instead of the original data. Since high-dimensional data for the features is necessary to acquire higher search accuracy, the metric space indices are indispensable to search fast for such high-dimensional data. The existing method using a *k*-nearest neighbor graph can reduce more computational cost than other previous works. On the *k*-nearest neighbor graph, each node is linked to the *k*-nearest neighbor nodes with edges, and then the proximity search is enabled by exploring from an arbitrary node to closer ones to a query object along the edges. However, there are two issues on the graph. First, instead of the small computational cost of the search, the huge computational cost is spent to construct the graph. Second, the search accuracy (recall rate) is reduced by disconnected graph nodes which cannot be traced.

In this study, an approximate *k*-nearest neighbor graph index with a tree-based index is proposed. The construction of the *k*-nearest neighbor graph spends the huge computational cost during the construction, because *k*-nearest neighbors should be found to add each node to the graph. Therefore, the proposed algorithm searches *k*-nearest neighbors using the partially constructed graph as a search index to reduce the computational cost and preserves the graph connectivity by adding nodes incrementally. Furthermore, closer nodes to a query object are searched using the tree-based index, then the closer nodes are used to explore the graph instead of an arbitrary node. On the other hand, some nodes on the graph tend to have a large number of edges which should be eliminated. Therefore, the algorithm checks whether the connectivity would be preserved or not by eliminating the excess edges. If the connectivity is preserved, the edges are eliminated. If the connectivity is not preserved, the edges are replaced to further nodes from the node having excess edges.

The cost of the proposed graph construction was reduced by about 96.7% compared to the *k*-nearest neighbor graph for 100,000 objects which are 50 dimensional uniform distribution data. The cost of the search was reduced by about 73% for 95% search accuracy compared to the *k*-nearest neighbor graph where the number of the edges is 16 for 100,000 objects which are 1,228 dimensional image feature data. This proposed algorithm reduced up to 34.2% edges on the graph for a million image feature objects. We concluded that the proposed algorithm can reduce the computational costs and the edges. Finally, the application example where the proposed algorithm was applied to a real internet product image search was introduced.