様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 4 月 15 日現在

機関番号：94305
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：20500109
研究課題名（和文）　複雑ネットワークを利用した大規模高次元データの高速類似探索技術

研究課題名（英文）　Fast similarity search methods for large-scale and high-dimensional data sets by using complex networks

研究代表者　上田　修功（UEDA NAONORI）
日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・所長

研究者番号：60379568

研究成果の概要（和文）：本研究では、スモールワールド特性を有する複雑ネットワーク（グラフ）である次数低減近傍グラフを索引構造とし、このグラフ上を貪欲探索又は最良優先探索アルゴリズムにより探索する類似探索法を提案した。提案法は、文書、画像、音声信号データ等の多様なメディアに適用でき、大規模高次元データに対しても高速類似探索を実現することを実験的に確認した。この性質を有する提案法は、新たな類似探索技術の礎となると期待できる。

研究成果の概要（英文）：This research project proposed a similarity search method that utilized a degree-reduced k-nearest neighbor graph (k-DR graph), which is a small-world network (or graph) in complex networks, for a search index structure, and that explored the k-DR graph by a greedy search or a best-first search algorithm. We have experimentally confirmed that this similarity search method was applicable to a variety of media types such as documents, images, and speech signals (utterances) and quickly found objects similar to a given query. The proposed similarity method with the foregoing properties is expected to serve as a base for novel similarity search techniques.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,500,000	450,000	1,950,000
２００９年度	1,300,000	390,000	1,690,000
２０１０年度	700,000	210,000	910,000
総　計	3,500,000	1,050,000	4,550,000

研究分野：機械学習
科研費の分科・細目：情報学、メディア情報学・データベース
キーワード：複雑ネットワーク、スモールワールド、探索、索引構造、アルゴリズム、大規模データ、ディレクトリ・情報検索、機械学習

１．研究開始当初の背景

文書、画像、音声音響信号データ等の多様な形態の情報が氾濫する情報爆発時代において、所望の情報を効率的に探索する技術は、情報処理の基盤となる必要不可欠な技術の 1 つである。

コンテンツ自体をクエリとし、クエリに類似するオブジェクトをデータベースから探す類似探索法は、ユークリッド空間中のベクトルとして表現されるデータを対象とした方法から、多様なデータ表現法を許す距離空間を対象とした方法へと、木構造を用いた方法を中心に改良されてきた [1], [2]。これらの方法は、一般的にクエリに対する厳密解を、距離公理を利用して効率的に求める方法である。木構造を利用した方法は、高次元データを対象とした場合、探索効率が著しく低下し、brute-force search（力まかせ探索）

と同程度の性能になるという欠点がある。

これらの厳密解を求める方法に対し、1990年代末頃から locality-sensitive hashing (LSH) [3] 等のハッシュ関数を用いた確率的方法による近似近傍探索の研究も進められてきた。特に、ハッシュ関数を用いた方法は、画像類似探索のように、画像自体を表す特徴やその特徴間の類似性が人為的に定められ、根本的に厳密なものではないデータの類似探索に、厳密性よりも効率性を重視し、適用されている。LSH は、対象とする空間や距離定義に関し強い制約を有し、メディア種や距離定義から独立した汎用的な方法ではない。

また、任意のオブジェクトの関係性を表現できるグラフを探索索引構造として利用する研究としては、単純な有向 k 近傍グラフ(有向 k-NN グラフ) を利用し、ヒューリスティックアルゴリズムで探索する研究が僅かになされている [4], [5]。

一方で、複雑ネットワークの研究は、スモールワールド特性を有するネットワークに関する Watts・Strogatz モデル (1998 年) [6] 端を発し、Kleinberg モデル (2000 年) [7] に代表されるモデル研究、社会現象や WWW 等のネットワークの現象解析への応用研究等が近年盛んである。しかし、スモールワールド特性を有するネットワーク (グラフ) を工学的に探索問題へ応用するというアプローチはなされていない。

[参考文献]

[1] C. Bohm, S. Berchtold, and D. A. Keim, “Searching in high-dimensional spaces - Index structures for improving the performance of multimedia databases,” ACM Computing Surveys, vol. 33, no. 3, pp. 322-373, 2001.
[2] G. R. Hjaltason and H. Samet, “Index-driven similarity search in metric spaces,” ACM Trans. Database Syst., vol. 28, no. 4, pp. 517-580, 2003.
[3] A. Andoni, et al, “Locality-sensitive hashing using stable distributions,” Nearest-neighbor methods in learning and vision: Theory and practice, MIT Press, Chapter 3, pp. 61-72, 2006.
[4] M. T. Orchard, “A fast nearest-neighbor search algorithm,” Proc. Int. Conf. Acoustics, Speech, Signal Process., vol. 4, pp. 2297-2300, 1991.
[5] T. B. Sebastian and B. B. Kimia, “Metric-based shape retrieval in large databases,” Proc. Int. Conf. Pattern Recognition, vol. 3, pp. 291-296, 2002.
[6] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” Nature, vol. 393, pp. 440-442, 1998.
[7] J. Kleinberg, “The small-world phenomenon: An algorithmic perspective,” Proc. ACM Symp. Theory of Computing (STOC), pp. 163-170, 2000.

２．研究の目的

本研究は、大規模な高次元マルチメディアデータを対象とした、超高速な類似探索法を確立し、類似探索法の新たな潮流を創造することを目的とする。目的達成の手段として、与えられたデータから探索用索引構造を事前に構築するアプローチを取り、索引構造として、多様なメディア種と非類似度定義を有する空間に適用可能なグラフ索引構造を用いる。特に、大規模高次元データを対象とした場合であっても高速類似探索を実現できる可能性のあるスモールワールド特性を有する複雑ネットワーク（スモールワールドグラフ）を利用する。この特性を有するグラフの 1 つである、高次元データから構築された k-近傍 (nearest neighbor：NN) グラフを基本的なグラフとする。

３．研究の方法

研究目的を達成するために、次の 3 つのプロジェクトを進める。

（１）複雑ネットワークを用いた基本探索法の有効性の検証。

① 近傍グラフにおけるスモールワールド特性の解析。

データの次元数とスモールワールド特性との関係に着目し、特に、高次元データから構築されたグラフが、類似度の高い頂点同士が結合されているにも係わらず、グラフの平均最短パス長が小さいというスモールワールド特性を有することを確認する。

② 探索アルゴリズムを効率的に動作させるグラフ構造の探求。

近傍グラフの中でも有望な次数低減近傍グラフを中心に検討を進める。

③ グラフ索引構造を用いた探索法の解の精度保証法の確立。

LSH 等のハッシュ関数を用いた近似近傍探索法を参考にし、確率的に精度を保証する近似アルゴリズムを模索するアプローチを取る。

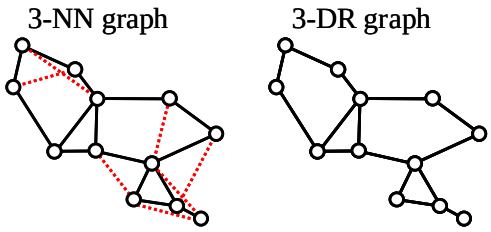
（２）大規模データの分散処理法の提案並びに有効性の検証。

与えられた大規模データを分割し、各分割データに対してグラフ索引構造を構築し、効率的探索を実現可能なように各索引構造を統合するアプローチを取る。データ分割の際に、LSH に代表されるハッシュ関数を利用する処理や統計学習技術によるデータ生成過程のモデル化等の方法の適用を検討する。

（３）マルチメディアデータに対する基本探索法の適用及び評価。
① 文書（テキスト）、画像、音声音響信号等の多様なメディア種や各メディア種に対して定義される非類似度を有する空間への基本探索法の適用と探索性能評価。
② 上記の単体メディア種を融合した複合メディアや複数の非類似度定義を融合した非類似度を有する空間への基本探索法の適用と探索性能評価。

これらの３つのプロジェクトの中でも、特に、基礎となる（１）に重点を置き研究を進める。

４．研究成果
（１）基本グラフである無向 k-NN グラフより、効率的な類似探索を実現するグラフである次数低減近傍グラフ（Degree-reduced k-nearest neighbor graph: k-DR graph）を見出し、構築アルゴリズムを開発した。k-DR グラフは、ある頂点 x とその近傍頂点 y との間の辺を、近傍頂点 y を起点とした貪欲探索アルゴリズムが頂点 x に到達できないときに限り、生成した近傍グラフである。このため、k-DR グラフは、各頂点の次数が、k-NN グラフと同じ又は小さいため、各頂点の近傍頂点から最良頂点を選択する計算量が同じ又は小さくなる。例として、12 頂点から構成される 3-NN グラフと 3-DR グラフとを次図に示す。



また、人工データ、実データの両方において、平均最短パス長は、k-NN、k-DR グラフ共に、ランダムグラフのものに近づく程度小さく、スモールワールド特性を有していた。特に、人工データを用いた次元数に対する平均最短パス長の解析では、次元数の増加に伴い、k-DR グラフの平均最短パス長は小さくなるという結果を得た。
探索精度保証問題に関しては、グラフ上のベイズンという尺度を用いて、貪欲探索成功確率を推定し、成功確率が与えられた値以上になることを保証する次数低減近傍グラフの構築法を提案した。画像データを対象として、この構築法及び探索法を評価し、有効性を確認した。

（２）統計的にオブジェクト間の近傍性を保持するハッシュ法の Locality-Sensitive Hashing（LSH）を実装し、分散型探索の前処理としての有用性を評価した。データの固有次元や非類似度定義に関する強い制約のため、既存の LSH を直接適用することは困難であると判断した。

（３）新聞記事等の文書データ、手書き文字や写真等の静止画データ、発話を混合ガウスモデル（GMM）で表現し、カルバックライブラーダイバージェンス（KLD）を非類似度とした非距離空間（GMM モデル空間）を探索対象データベースとした場合、k-DR グラフを索引構造とした探索法は既存法に対して優れた高速探索性能を示した。特に、GMM モデル空間に代表される一般的な非距離空間を扱うことができる汎用類似探索法は他に類を見ない方法である。
画像データに関して、色ヒストグラムと方向性特徴（SIFT）との異なる２つの特徴に基づく非類似度を同時に用い、k-DR グラフ索引構造を用いた探索法を評価し、良好な結果を得た。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）
① 青山一生、斉藤和巳、山田武士、上田修功、グラフ索引構造を用いた高速類似探索電子情報通信学会論文誌 D、J92-D、1543-1554、2009、査読有
② 木村学、斉藤和巳、上田修功、効率的な類似探索のためのピボット学習法、情報処理学会論文誌、50、1883-1891、2009、査読有
③ K. Saito, T. Yamada, K. Kazama, “Extracting communities from complex networks by the K-Dense method,” IEICE Trans. Fundamentals, E91-A, 3304-3311, 2008, 査読有

〔学会発表〕（計５件）
① 外岡達也、小出明弘、斉藤和巳、青山一生、澤田宏、上田修功、ネットワーク構造による類似探索性能の分析法の提案、第9回情報科学技術フォーラム（FIT2010）、2010年9月8日、九州大学　福岡市
② 小出明弘、外岡達也、斉藤和巳、青山一生、澤田宏、上田修功、オブジェクト集合に依存したＲＮＧの特性分析、第９回情報科学技術フォーラム(FIT2010)、2010 年９月８日、九州大学　福岡市
③ K. Aoyama, S. Watanabe, H. Sawada, Y. Minami, N. Ueda, K, Saito, “Fast similarity search on a large speech data set with neighborhood graph indexing,” IEEE Int. Conf. Acoustics, Speech and

Signal Processing (ICASSP2010), 2010年3月19日, Dallas, Texas, USA
④ K. Aoyama, K. Saito, T. Yamada, N. Ueda, “Fast similarity search in small-world networks,” Int. Workshop on Complex Networks (ComplNet2009), 2009年5月26日, Catania, Italy
⑤ 青山一生、斉藤和巳、山田武士、上田修功、ネットワーク索引構造を用いた類似探索と可視化、第7回情報科学技術フォーラム、2008年9月4日、慶應義塾大学、湘南藤沢

〔図書〕(計1件)
① K. Aoyama, K. Saito, T. Yamada, N. Ueda, “Fast similarity search in small-world networks,” in Complex Networks, Studies in Computational Intelligence, Springer, pp. 185-196, 2009

６．研究組織
(1)研究代表者
上田　修功 (UEDA NAONORI)
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・所長
研究者番号：60379568

(2)研究分担者
斉藤　和巳 (SAITO KAZUMI)
静岡県立大学・経営情報学部・教授
研究者番号：80379544

山田　武士 (YAMADA TAKESHI)
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・協創情報研究部・主幹研究員
研究者番号：50396115
(2009年度に研究分担者を辞退)

澤田　宏 (SAWADA HIROSHI)
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・協創情報研究部・主幹研究員
研究者番号：10396210
(2009年度に研究分担者として加入)

青山　一生 (AOYAMA KAZUO)
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・協創情報研究部・主任研究員
研究者番号：80447028