D-040

ネットワーク索引構造を用いた類似探索と可視化

Search and Visualize Similar Objects by Utilizing Networks as Index Structures

青山 一生† 斉藤 和巳‡ 山田 武士† 上田 修功†
Kazuo Aoyama Kazumi Saito Takeshi Yamada Naonori Ueda

1. まえがき

高次元のオブジェクト集合から, 与えられたクエリに類似するオブジェクトを高速に探索する類似探索を考える. 探索性能の向上に加えて, 探索結果を 2 次元又は 3 次元空間に配置し, 可視化することは, ユーザの利便性を更に向上する. このような類似探索とその結果の可視化に用いられる手法として, FastMap が挙げられる [1]. FastMap では, オブジェクト集合を次元縮約し低次元空間に埋め込み, その低次元空間において探索を実行し, 可視化を行う. しかしながら, このような方法では, 視認性の良いオブジェクト配置が必ずしも探索性能の向上に結び付かないという問題がある.

本稿では, 索引構造として, 効率的な探索を実現するように設計した近傍ネットワークを用い, 類似探索を, ネットワークのリンクを順次辿ることで類似ノードを見つけるネットワーク近傍探索問題と捉える新たなアプローチを提案する. また, 探索結果のネットワークは索引構造ネットワークの部分グラフであるため, 探索性能を損なうことなく, 視認性の高い探索結果の可視化を実現できる.

2. 問題設定

高次元オブジェクト集合 $\mathcal{X}$, 探索対象オブジェクト集合 $X \subset \mathcal{X}$, クエリ集合 $X_q \subset \mathcal{X}$ とし, $x, y \in \mathcal{X}$ に関し, 距離 $d(x,y)$ 又は類似度 $\rho(x,y)$ が定義されていると仮定する. 本稿では, クエリ $q \in X_q$ に最類似のオブジェクトの 1 つ $x^* \in X$ を見つける探索問題を対象とする. この問題を, X から構成された無向近傍ネットワーク Γ （Γ の構成アルゴリズムは後述）, オブジェクト x を Γ のノード x, 起点ノードを $x_0 \in X$ とし, Γ のリンクを辿りながら x^* を見つけるネットワーク探索問題として考える.

3. 提案方法

3.1 探索アルゴリズム及び探索性能指標

与えられた Γ のリンクを辿りながら順次探索するアルゴリズムとして, 最良優先近傍探索（best-first neighborhood search: BS）アルゴリズムを用いる. BS アルゴリズムは, Γ, 探索コストの上限値 β（詳細は後述）, q, 起点ノード x_0 を受け, 最類似ノードを返す. 探索過程での最類似ノードを z とし, $z \leftarrow x_0$ で初期化する. z と q との類似度 $\rho(z,q)$, 及び, z に直接リンク接続された近傍ノード集合 $\Gamma(z)$ と q との類似度を求める. ここで, q との類似度計算済ノード集合を A とし, $B = \{x \in X; \Gamma(x) \in A\}$ とする. 次に, $A \setminus B$ のノード y に関し, $x = \arg\max_{y \in A \setminus B}\{\rho(y,q)\}$ を求め, $\rho(z,q) < \rho(x,q)$ ならば, $z \leftarrow x$ で z を更新し, $A \leftarrow A \cup \Gamma(x)$, $B \leftarrow B \cup \{x\}$ とする. 再び, $A \setminus B$ を求め, 手続きを繰り返す.

実験では, 式 (1) の探索コスト期待値 $C(\Gamma)$ と式 (2) の成功率 $S(\Gamma, \beta)$ とを用い, 探索性能を評価する. 但し, 予め q に対する最類似オブジェクト x^* を求め, x^* を発見した時点又は β に達した時点で探索終了とした. また, 各クエリに対する起点ノード集合を X_0, 探索終了時の A のノード数を $|A|$ とした.

$$C(\Gamma) = \frac{1}{|X_q||X_0|} \sum_{q \in X_q} \sum_{x_0 \in X_0} |A| . \quad (1)$$

$$S(\Gamma;\beta) = \frac{1}{|X_q||X_0|} \sum_{q \in X_q} \sum_{x_0 \in X_0} \delta(q, x_0; \Gamma, \beta) , \quad (2)$$

$$\delta(q, x_0; \Gamma, \beta) = \begin{cases} 1 & \text{for} \quad z = x^* \\ 0 & \text{for} \quad z \neq x^* \end{cases} . \quad (3)$$

3.2 ネットワーク構成アルゴリズム

オブジェクト x に類似する上位 k 個のオブジェクト集合を $N_k(x)$ とし, x の近傍ノード集合 $\Gamma_k(x)$ を $N_k(x) \cup \{y \in X; x \in N_k(y)\}$ と定義する. $\Gamma_k(x)$ は, x を端点とするリンクも規定しているため, 全ての $x \in X$ に対する $\Gamma_k(x)$ により, X をノード集合とするネットワーク Γ_k を規定できる. この Γ_k を NN ネットワーク Γ_{kNN} と呼ぶ.

Γ_{kNN} を基本構造とし, BS アルゴリズムによる探索性能を向上した次数低減（degree-reduced: DR）NN ネットワーク Γ_{DR} を構成する次数低減（degree-reduction: DR）アルゴリズムを提案する. DRNN ネットワークに BS アルゴリズムを適用し探索する提案方法を DR 探索方法と呼び, 同様に, ベースラインである NN ネットワークを用いた方法を NN 探索方法と呼ぶ.

DR アルゴリズムは, $\Gamma_{k'DR}$ を $\Gamma_{(k'-1)DR}$ から構成する逐次的アルゴリズムである. オブジェクト x に対し, $\{y\} = N_{k'}(x) - N_{k'-1}(x)$ となる y と x との間を直接リンク接続するか否かを考える. y から x に, 貪欲探索（greedy search: GS）アルゴリズム $GS(x, y; \Gamma_{(k'-1)DR})$ により, 到達可能でない（FALSE を返す）ときに限り, DR アルゴリズムは直接リンクを生成する. GS アルゴリズムは, y から, x に対する類似度が増加するように順次リンクを辿り, x に到達可能か否かを判定するアルゴリズムである. このように, 間接的に接続されているオブジェクト間に新たなリンクを生成しないことで平均次数を低減する. DR アルゴリズムを次頁に具体的に示す.

3.3 探索結果可視化

Γ_{kDR} を用いて探索した結果は, クエリに類似する無向近傍ネットワークとして得られる. このため, バネモデル [3] のような小規模グラフを高速に可視化するアルゴリズムを利用することができる.

†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
‡静岡県立大学

Algorithm 1 *DR Algorithm* ($k \geq 2$)

Input: X, Γ_{1DR} (=Γ_{1NN})
Output: Γ_{kDR}

1. **for** $k' = 2$ to k **do**
2. $\Gamma(x)_{k'DR} = \Gamma(x)_{(k'-1)DR}$ for all $x \in X$
3. Lines 4–7 are executed for all $x \in X$.
4. $\{y\} = N_{k'}(x) \setminus N_{k'-1}(x)$
5. **if** $GS(x, y; \Gamma_{(k'-1)DR})$ == **FALSE: then**
6. $\Gamma(x)_{k'DR} = \Gamma(x)_{k'DR} \cup \{y\}$
   $\Gamma(y)_{k'DR} = \Gamma(y)_{k'DR} \cup \{x\}$
7. **end if**
8. $k' = k' + 1$
9. **end for**

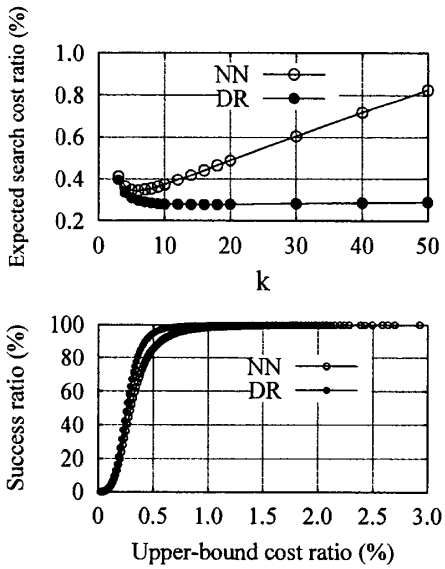

図 1: 探索結果, 上図: k に対する探索コスト, 下図: 探索コスト上限値率に対する成功率

4. 実験

4.1 データ及び実験方法

手書き文字データベース MNIST [4] を $\mathcal{X}$, 60,000 オブジェクト（0〜9 の数字）を X, 10,000 オブジェクトを X_q（但し, $X \cap X_q = \emptyset$）とした. 各オブジェクトは, 28×28 ピクセルの各々を 256 階調グレースケールで数値化した 784 次元ベクトルで表されている. 本実験では, 正規化ベクトルを特徴量とし, ユークリッド距離を用いた. 各クエリに対し 10 回の探索を行い, 探索コスト期待値と成功率とを求めた. 各探索について, X から一様ランダム選択したノードを起点ノードとした.

4.2 結果

図 1 上は, NN, DR 探索方法の探索コスト期待値率（探索コスト期待値の全オブジェクト数 $|X|$ に対する比率 %）の k 依存性を示す. DR 方法は, $k = 16$ において 0.28% を達成し, 高効率な探索を実現している. 図 1 下は, 図 1 上で最小探索コスト期待値率を達成した k を用いたネットワークを索引構造とした場合の, 探索コスト上限値率（探索コスト上限値の $|X|$ に対する比率 %）に対する成功率を示す. DR 方法は, 上限値率を高々0.43% に設定した場合, 90% の成功率で最類似オブジェクトを見つけることができた. 次に, 探索結果の近傍ネットワークをバネモデルを用いて可視化した例を示す. 図 2 は, クエリとして「5」を用い, 類似する上位 50 を探索し, 得られた最大コンポーネントのネットワーク（ノード数 45）の可視化例である. この例では, クエリである「5」に類似する「3」,「8」との関係を把握することができる.

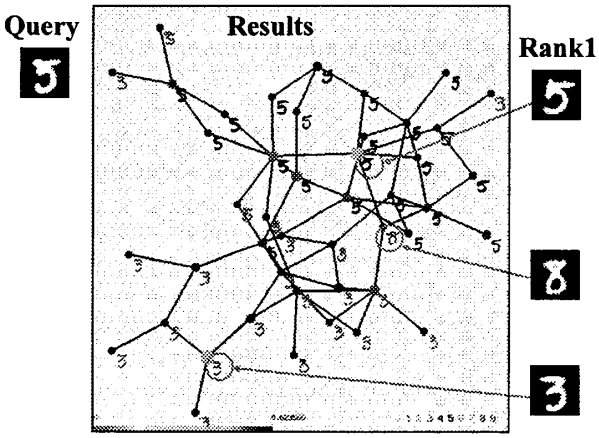

図 2: 類似探索結果可視化例

5. むすび

ネットワークを索引構造に用いた高速類似探索及び可視化方法を提案した. 索引構造である次数低減近傍（DRNN）ネットワークは, 新たに提案した次数低減アルゴリズムにより, 高次元オブジェクト集合から直接構成された. この DRNN ネットワークに, 最良優先探索アルゴリズムを適用した提案探索方法は, MNIST を対象とした実験において, 非常に優れた探索性能を示した. 更に, 探索結果は小規模なネットワークであるため, バネモデルを用いて高速に描画できた.

謝辞

本実験の補助に関し, 藤本裕文氏に感謝致します. 尚, 本研究の一部は, 日本学術振興会科学研究費補助金 基盤研究 (C)（課題番号：20500109）を受けて行われた.

参考文献

[1] C. Faloutsos and K.-I. Lin, FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets, Proc. ACM SIGMOD, pp. 163–174, 1995.

[2] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín, Searching in metric spaces, ACM Comp. Surveys, vol. 33, no. 3, pp. 273–321, 2001.

[3] T. Kamada and S. Kawai, An algorithm for drawing general undirected graphs, Information Processing Lett., vol. 31, pp. 7–15, 1989.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.