

複数の非類似度を自由に重みづけできる 単一のグラフ索引を用いた最近傍探索手法

田村 真一^{1,a)} 澤田 宏^{2,b)} 山田 武士^{3,c)}

概要: 本稿では、探索ごとに複数の非類似度の重みづけを自由に変えて探索できる、単一の索引を用いた最近傍探索手法を提案する。多くの最近傍探索問題において非類似度の基準は一つとは言えず、それらの中でどの非類似度をどの程度重視するかを探索のたびに設定できる手法が求められている。これまでに提案されている手法で複数の非類似度を取り扱うには、事前に重みづけを固定するか、それぞれの非類似度での結果を統合する方法がとられることが多かった。しかし事前に重みを固定する手法では探索時に重みを調整することは難しく、また結果統合の手法ではそれぞれの基準では中位に来るが重みつけた基準では上位に来るような結果が出力されにくい。本研究では、複数の非類似度に対する任意の結合重みを考慮した単一のグラフを索引に用いることで、探索時にはその結合重みを自由に設定できる最近傍探索手法を提案する。実画像データセットを使ったさまざまな重みづけに対する実験の結果、提案手法は索引を1回しか構築しないにもかかわらず、それぞれの重みで固定して作った従来の索引構造に匹敵する探索コストと精度を達成することができた。

Nearest neighbor search method under variably-weighted multiple dissimilarities using a single graph index

SHINICHI TAMURA^{1,a)} HIROSHI SAWADA^{2,b)} TAKESHI YAMADA^{3,c)}

Abstract: This paper introduces the nearest neighbor search method using a single index, which allows users to put weights on the multiple dissimilarities for each search trial. Many of nearest neighbor search problems come up with more than one dissimilarity, which are desired to be weighted on-the-fly. To cope with this sort of multiple dissimilarities using existing search methods, we usually fix the weights beforehand or aggregate results from each dissimilarity afterward. However, pre-fixed weights are not adjustable on the search stage, and result aggregation methods are unlikely to give a result which would be top-ranked on the weighted dissimilarity while middle on each of the unweighted dissimilarities. We propose the novel nearest neighbor search method which allows users to set combination weight of multiple dissimilarities at the search stage, by using a single graph index constructed taking arbitrary weights into consideration. In the experiment with the real image dataset and variable weights, our index performs comparably to the conventional indices which is constructed with pre-fixed weights although our index is constructed only once.

1. はじめに

ウェブメディアの発達とともに日々大量のデータと接するようになった今日、あるアイテムに似たアイテムを膨大なデータベースから探し出す技術が必要な場面が増えている。たとえば、利用者が撮った写真と似ている写真を画像共有サイトから探したいときや、利用者が購入した商品に近いものを商品データベースから探したいときなどである。

¹ 奈良先端科学技術大学院大学

Nara Institute of Science and Technology

² NTT サービスエボリューション研究所

NTT Service Evolution Laboratories

³ NTT コミュニケーション科学基礎研究所

NTT Communication Science Laboratories

a) tamura.shinichi.tl1@is.naist.jp

b) sawada.hiroshi@lab.ntt.co.jp

c) yamada.tak@lab.ntt.co.jp

最近傍探索問題とは上記のように、キーワードではなくアイテムをクエリとして、それに最も類似したアイテムをデータベースから探す問題である。このとき、データベースは大規模で検索は多数回行われるのが一般的である。そのためクエリアイテムをデータベース内の全アイテムと総当りで比較するのは実用に適さず、検索に先立って索引を作成しておき1回1回の検索を効率化する必要がある。

そのような索引として、LSH (局所性鋭敏型ハッシュ) [1], 木構造 [2], [3], [4], [5], [6], 木構造以外のグラフ構造 [7], [8], [9], [10] などを用いた手法が挙げられる。LSHは、類似アイテム同士が衝突しやすいハッシュ関数を用いることで高次元データに対しても高速に探索できる。しかし L^2 ノルム (ユークリッド距離) [11] やコサイン類似度 [12] など特定の類似度以外には適切なハッシュ関数が提案されていない。木構造を用いる手法としては各次元にそって空間を分割する kd 木が最もよく知られている。 kd 木は厳密探索が可能であるかわりに高次元データに対しては総当りと同程度の探索コストがかかる。またグラフ構造を用いる手法としてポロノイ分割の双対であるドロネーグラフ [7] が広く知られている。やはり厳密探索が可能かわりにユークリッド距離以外の距離空間や高次元空間においてはグラフの生成が難しく実用的でない。

以上の手法はいずれも基本となるアイデアは空間分割に基づく手法であり、どのような類似度関数を用いるかによって索引の構築の仕方が変わってくる。そのため、アルゴリズムが対応していない類似度関数を使う場合や、アイテムごとにすでに計算された類似度の値しか得られない状況では利用できない。いっぽう本研究で提案するグラフ索引は、アイテムごとの類似度の値さえ得られていれば適用可能な手法である。そのような手法としては、 K 最近傍グラフ [9] や、 K 最近傍グラフの探索コストを大幅に削減した次数低減 K 最近傍グラフが提案されている [10]。

ところである2つのアイテムが与えられたとき、それらの類似度をどういった特徴に着目して測るべきかは、利用者やシチュエーションによって変わり得るため事前に一意に定めることは難しい。場合によっては、ある特徴で測った類似度を重視しつつ、別の特徴による類似度も加味して検索をしたいというニーズもあるだろう。

たとえば画像であれば、局所的なテクスチャによる Bag-of-Visual-Words [13] といった特徴量もあれば大域的な色に基づくカラーヒストグラムといった特徴量、あるいは撮影者・撮影日時やカメラの設定などのメタ情報に基づく特徴量があり、それぞれで測る類似度は全く異なる。同じものが写っている画像を得るためには Bag-of-Visual-Words のみをもとに探索するのが適当だとしても、写っているものが多少異なっても同じ質感の写真が欲しい場合にはこれに加えてカラーヒストグラムやレンズの絞り値などの情報も考慮に入れたほうが良いといったことも考えられる。

より実用的な応用例として、EC サイトの商品検索やレコメンデーション、レストラン・不動産検索といった応用も挙げられる。この場合の類似度としては値段や各種スペックなど様々であり、どれか一つだけの基準ではなく利用者の複数の関心をバランスよく満足する結果が望まれる。

従来はこういった複数の類似度を扱う問題に対して、索引を必要としない程度の小規模な問題であれば、事前に特徴量を複合することで対応できた [14]。しかし索引が必要となるほどの大規模な問題の場合、事前にあらゆる複合のしかたで特徴量を複合してそれぞれに応じた索引をすべて作っておくことは難しいため、個別の特徴量ごとに索引を作って探索を行ってからその結果を統合する方法がとられることが多い [15]。だがこういった結果統合の手法においては、最終的に出力されるのは基本的にはいずれか単一の基準で上位に来る結果であり、重み付けされた基準では上位に来るが個々の基準では上位ではなく中位になってしまう (つまり複数の基準を反映している) 結果は出力されにくい。複数の類似度を取り扱い、かつ大規模な最近傍探索問題に直接取り組んだ手法は未だ提案されていない。本研究では、この問題に対して有効なグラフ構造を用いた索引を提案する。

2. 問題設定

2.1 単一非類似度の最近傍探索問題

まずは単一の非類似度を扱う最近傍問題を定式化する。

最近傍探索問題

データベースを X , 非類似度を d とする。
クエリアイテム q が与えられたときに、下記を満たすアイテム x^* を探せ:

$$x^* = \arg \min_{x \in X} d(q, x).$$

なお定式化にあたっては、似ているほど小さい値をとる非類似度を用いた。これは必ずしも距離関数である必要はなく、類似度の符号を反転する、あるいは最大値からの差を取るなどの方法で得られる。

このような問題が与えられた時、下図1に示したように事前に索引を構築することで、大規模なデータベースに対しても1回1回の探索を高速化することができる。

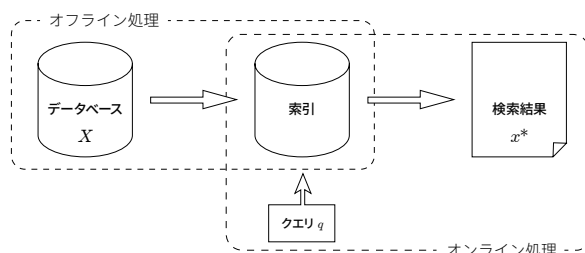


図1 索引を用いた最近傍探索システム

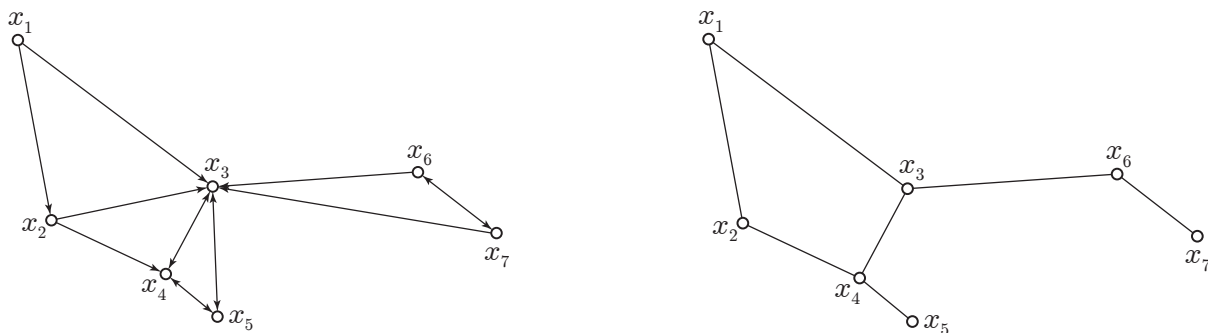


図 2 2次元平面における 2-NN (左) および 2-DR グラフ (右) の例

ここで、あるデータ点の近傍にどのデータ点があるかという局所的な情報があれば、連続最適化における最急降下法同様に非類似度を減少する方向へデータベース内を降下していくことができる。この近傍情報を重みなしのグラフとして表現したのがグラフ索引である。

グラフ索引が与えられたとき、探索は下記のリスト 1 のように、貪欲法によって辺をたどって頂点間を遷移して行う。ただしグラフ索引では木と異なり任意の頂点から遷移を開始できるので、探索精度を上げるために、初期頂点を変えて L 回 (あるいは L 並列で) 探索試行を行うことができる。なお、 $\text{Adj}(x)$ は x の隣接頂点を表す。

リスト 1: グラフ索引を用いた最近傍探索

```

1 for  $\ell \leftarrow 1 \dots L$  do
2    $x_\ell^+ \leftarrow$  初期頂点; // 一般にランダムに設定
3   do
4      $x_\ell \leftarrow x_\ell^+$ ;
5      $x_\ell^+ \leftarrow \arg \min_{a \in \text{Adj}(x_\ell)} d(q, a)$ ;
6     while  $d(q, x_\ell^+) \leq d(q, x_\ell)$ ;
7   end
8    $x^* \leftarrow \{x_\ell\}$  の中で最も  $q$  に近いもの;
9 return  $x^*$ 

```

以下、前節で紹介した手法の中から本研究と関わりのあるものについて詳説する。ここで、簡単のため x の K 最近傍リストを $K\text{-NN}_x$ と書き、そのうち k 番目の要素 (つまり x から k 番目に近い要素) を $K\text{-NN}_x(k)$, i 番目から j 番目までの要素を $K\text{-NN}_x(i:j)$ と書く。

2.1.1 K -最近傍グラフ ($K\text{-NN}$ グラフ)

Sebastian らは、高次元空間にて構築の難しいドロネーグラフに代わって K 最近傍グラフ ($K\text{-NN}$ グラフ) を用いた最近傍探索手法を提案した [9]。 K 最近傍グラフとは、図 2 に例示したような、各頂点 x から $K\text{-NN}_x$ へ辺を張った有向グラフのことである。なお Sebastian らの提案手法では、リスト 1 の 6 行目において隣接頂点より広い範囲の頂点を探索するが、ここではその詳細までは述べない。

2.1.2 次数低減 K -最近傍グラフ ($K\text{-DR}$ グラフ)

$K\text{-NN}$ グラフにおいては、探索精度と計算量はパラメータ K のもとでトレードオフ関係にある。そこで青山らは、

計算量を抑えたまま精度を向上させたグラフ索引として次数低減 K 最近傍グラフ ($K\text{-DR}$ グラフ) を提案した [10]。図 2 にある例からも分かる通り、これは $K\text{-NN}$ グラフのうち探索に不要な辺を削減し辺を無向にしたグラフであり、リスト 2 のように構築する。

リスト 2: $K\text{-DR}$ グラフの構築

```

1 for  $k \leftarrow 1 \dots K$  do
2   for 各頂点  $x$  について do
3      $y \leftarrow K\text{-NN}_x(k)$ ;
4     if  $y$  から  $x$  が貪欲探索可能でないなら then
5        $x$  から  $y$  へ無向辺を張る;
6     end
7   end
8 end
9 end

```

貪欲探索可能であるとは以下のように定義される。

定義 1 貪欲探索可能: 頂点 y から頂点 x が貪欲探索可能であるとは、クエリを x , 初期頂点を y としてリスト 1 によって探索した際に x が解になること。

グラフ構築にあたっては、ある頂点から別の頂点へ貪欲探索可能かどうか素早く判定することが重要になる。これは、 $K\text{-DR}$ グラフでは以下の定理 1 により判定できる。

定理 1: $y = K\text{-NN}_x(k)$ から x への辺を張るか検討する時点 (リスト 2 の 5 行目) で、 y と $\exists z \in K\text{-NN}_x(1:k-1)$ の間に辺があるとき、かつそのときに限り、 y から x は貪欲探索可能である。

この定理を証明するために、保証領域と遷移領域という 2 つの領域を定義する。

定義 2 保証領域 \mathcal{S}_x : データベース X 中の点で、すでに x との間に辺を張るか検討した点の集合。

このとき、後の段階で追加された辺が既に保証領域に入っていた点の貪欲探索可能性を破壊しないため、下記の補題が成立する。

補題 1: $\forall z \in \mathcal{S}_x$ から x へは常に貪欲探索可能である。

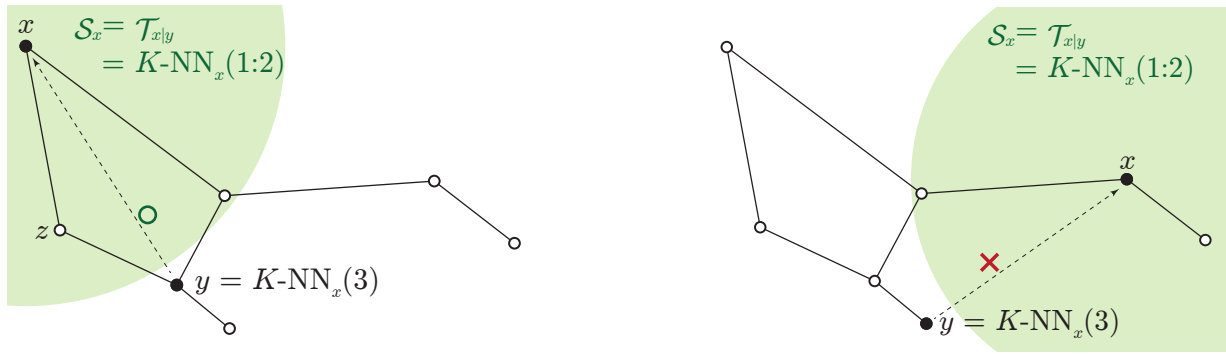


図3 定理1の例. (左) y から $z \in K\text{-NN}_x(1:2)$ に辺があるため x まで貪欲探索可能.
(右) y から $K\text{-NN}_x(1:2)$ 内の点に辺がないため x まで貪欲探索可能ではない.

定義3 遷移領域 $\mathcal{T}_{x|y}$: y に比べ, より x に近い点の集合.

$$\mathcal{T}_{x|y} = \{t \in X \setminus \{y\} \mid d(x, t) \leq d(x, y)\}.$$

x をクエリとして探索する際, $t \in \mathcal{T}_{x|y}$ と y の間に辺があれば, y の次に t に遷移する可能性がある.

遷移領域の定義により, $y = K\text{-NN}_x(k)$ に対して $\mathcal{T}_{x|y} = K\text{-NN}_x(1:k-1)$ である. いっぽう $K\text{-DR}$ グラフの構成にあたっては x に近い頂点から順に辺を張っていくため, $\mathcal{S}_x = K\text{-NN}_x(1:k-1)$ でもある. すなわち,

補題2: $y = K\text{-NN}_x(k)$ から x への辺を処理する時点で, $\mathcal{S}_x = \mathcal{T}_{x|y} = K\text{-NN}_x(1:k-1)$ である (下図4参照).

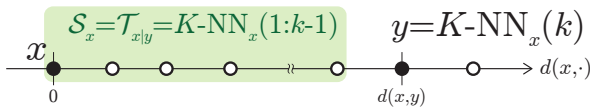


図4 $d(x, \cdot)$ によって各頂点を数直線上に配した図.
 $\mathcal{S}_x = \mathcal{T}_{x|y} = K\text{-NN}_x(1:k-1)$ となる (補題2).

(定理1の証明) 補題1, 2より, $y = K\text{-NN}_x(k)$ から x への辺を処理する時点で $K\text{-NN}_x(1:k-1)$ から x へは貪欲探索可能である. よって, y と $\exists z \in K\text{-NN}_x(1:k-1)$ の間に辺がある場合 (たとえば図3左) には y から z へ遷移することで x に近づき, さらに z は保証領域 \mathcal{S}_x 内にあるので x へ到達することができるため, 結局 y から x も貪欲探索可能であることになる. 逆に辺がない場合 (たとえば図3右) には, y から x に近づくことが出来ずこれ以上遷移することはできない. 以上により, 定理1が示された.

2.2 複数非類似度の最近傍探索問題

本研究では, 上記の最近傍探索問題において非類似度が複数ある場合を取り扱う. この問題において, 利用者はその複数ある非類似度の中から一つ以上を選び, さらにそれらを任意の重み付けで線形結合して探索できるものとする.

簡単のため本稿では非類似度は2つとする. 3つ以上の場合への拡張は今後の課題である.

2つの非類似度のもとでの最近傍探索問題

データベースを X , 2つの非類似度を d^1, d^2 とする. クエリアイテム q と非類似度の重視パラメータ $w \in [0, 1]$ が与えられたときに, 下記を満たすアイテム x^* を探せ:

$$x^* = \arg \min_{x \in X} d(q, x; w),$$

ただし,

$$d(q, x; w) = wd^1(q, x) + (1-w)d^2(q, x).$$

本稿では上記のような問題に対して有効な索引を提案する. ただし下図5にある通りパラメータ w が与えられるのは検索時点であり, 索引作成時点では X, d^1, d^2 のみしか与えられない. この状態で任意のクエリアイテム q はもちろん, 任意の重視パラメータ w に対して有効な索引を作成することを目標とする.

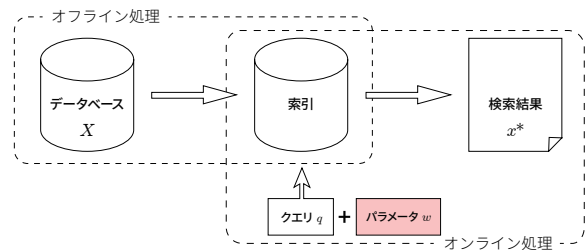


図5 パラメータを含む場合の検索システム
(パラメータが与えられるのはオンライン処理の段階である)

3. 提案手法

本稿では $K\text{-DR}$ グラフを複数非類似度へ拡張した多モード K 最近傍グラフ ($K\text{-MN}$ グラフ) を提案する. これは $K\text{-DR}$ グラフ同様, $k = 1 \dots K$ で $k\text{-NN}$ グラフを作るが探索に不要と判断された辺は張らない. ただし索引を任意の重視パラメータ w に対応させるには, 本来は索引構築に用いる K 最近傍リストも任意の w に対するものを用いるべきである. しかし w は連続値なので不可能であるため, かわりに各非類似度単体での最近傍リスト ($w = 0$ と $w = 1$

の場合に相当)のみを用いる。K-MN グラフの具体的な構築アルゴリズムは下記リスト 3 の通り。なお、 $K\text{-NN}_x^d$ は非類似度 d における頂点 x の K 最近傍リストを表す。

リスト 3: K-MN グラフの構築

```

1 for  $k \leftarrow 1 \dots K$  do
2   for 各頂点  $x$  について do
3     for 各非類似度  $d$  について do
4        $y \leftarrow K\text{-NN}_x^d(k)$ ;
5       if  $y$  から  $x$  が  $\exists w$  に対し貪欲探索可能でないなら then
6          $x$  から  $y$  へ無向辺を張る;
7       end
8     end
9   end
10 end

```

K-DR グラフと違い K-MN グラフにおいては探索時に重視パラメータ w が与えられる。索引構築時にはこの w としてあらゆる値に対して y から x が貪欲探索可能であることを検証しなくてはならない。

しかし残念ながら定理 1 は成立しないため、前出の定義・補題を改めた上で別の判定基準を設ける必要がある。

定義 3' 遷移領域 $\mathcal{T}_{x|y}$: y に比べ、いずれかの基準でより x に近い点の集合。

$$\mathcal{T}_{x|y} = \{t \in X \setminus \{y\} \mid d^1(x,t) \leq d^1(x,y) \vee d^2(x,t) \leq d^2(x,y)\}.$$

x をクエリとして探索する際、 $t \in \mathcal{T}_{x|y}$ と y の間に辺があれば、 $\exists w$ に対しては y の次に t に遷移する可能性がある。

補題 2 も成立せず、遷移領域 $\mathcal{T}_{x|y}$ のうち保証領域 \mathcal{S}_x に含まれない領域が存在する。これを危険領域 $\mathcal{D}_{x|y}$ と呼ぶ。

定義 4 危険領域 $\mathcal{D}_{x|y}$: 遷移領域 $\mathcal{T}_{x|y}$ のうち、保証領域 \mathcal{S}_x に含まれない頂点の集合。

$$\mathcal{D}_{x|y} = \mathcal{T}_{x|y} \setminus \mathcal{S}_x.$$

図 6 は 2 次元平面 $d^2(x, \cdot) - d^1(x, \cdot)$ 上に示した保証領域 \mathcal{S}_x と危険領域 $\mathcal{D}_{x|y}$ の例である。この例において y は d^1 の基準においては k 番目であるが、 d^2 での順位はそれよりも後であるため y と保証領域 \mathcal{S}_x の間に危険領域 $\mathcal{D}_{x|y}$ が存在する。もし y から $z \in \mathcal{D}_{x|y}$ への辺が存在する場合、パラメータ w の値によっては (たとえば $w = 0$ など) z の方がより x に近いので、 y から z へ遷移してしまう可能性がある。この z から x への貪欲探索可能性はこの時点では検討されていないため、 y から x も貪欲探索可能とは言えない。

なお補題 1 も厳密には成立せず、追加された辺がそれ以前に保証されていた別の頂点对の貪欲探索可能性を破壊することがあるが、4 節で後述する通りこれは実用上無視できる程度である。

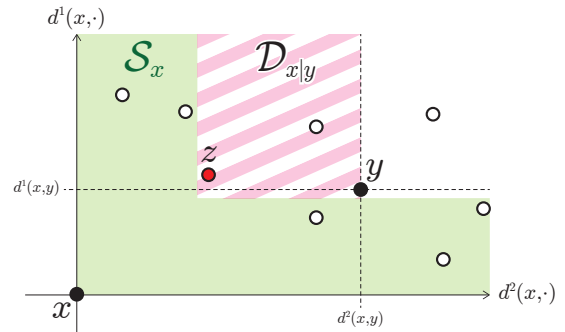


図 6 保証領域 \mathcal{S}_x と危険領域 $\mathcal{D}_{x|y}$ の例

以上を踏まえ、貪欲探索可能性の判定は下記のように行う。まず、 y と \mathcal{S}_x 内の頂点との間に辺がない場合、K-DR 同様に貪欲探索可能でない。また、 y と $\mathcal{D}_{x|y}$ 内の頂点との間に辺がある場合も、やはり貪欲探索可能でない。これらの条件は貪欲探索可能でないことの必要条件ではないが、いずれにも当てはまらない場合は貪欲探索可能であると判定する。この判定法を含め、リスト 3 を再掲する。

リスト 3': K-MN グラフの構築 (詳細)

```

1 for  $k \leftarrow 1 \dots K$  do
2   for 各頂点  $x$  について do
3     for 各非類似度  $d$  について do
4        $y \leftarrow K\text{-NN}_x^d(k)$ ;
5       if  $y$  から  $\forall z \in \mathcal{S}_x$  に辺がない, または
6          $y$  から  $\exists z' \in \mathcal{D}_{x|y}$  に辺があるなら then
7          $x$  から  $y$  へ無向辺を張る;
8       end
9     end
10  end

```

4. 実験

4.1 Holidays データセット

今回の実験では、実際の画像データセットとして Holidays データセット [16] を採用した。これは個人の休暇の写真を中心に集められた 1491 枚の画像からなるデータセットであり、様々なシーンの写真を含んでいる。

また非類似度を測る複数の特徴量として、 $H \times S \times V = 20 \times 10 \times 10 = 2000$ 次元のカラーヒストグラムおよびクラス数 1000 の Bag-of-Visual-Words の 2 種類の特徴を抽出し、それぞれの特徴量に対応する非類似度関数としてコサイン非類似度 $d_{\cos}(x, y) = 1 - x^T y / \|x\| \|y\|$ を用いた。

提案法の必要性を確認するため、実際にこのデータセットに対して、両非類似度間の重みを変えながら全探索する予備実験を行った。得られた厳密解の例を最終頁の図 8 に 2 つ示す。1 つ目の例では、Bag-of-Visual-Words のみを

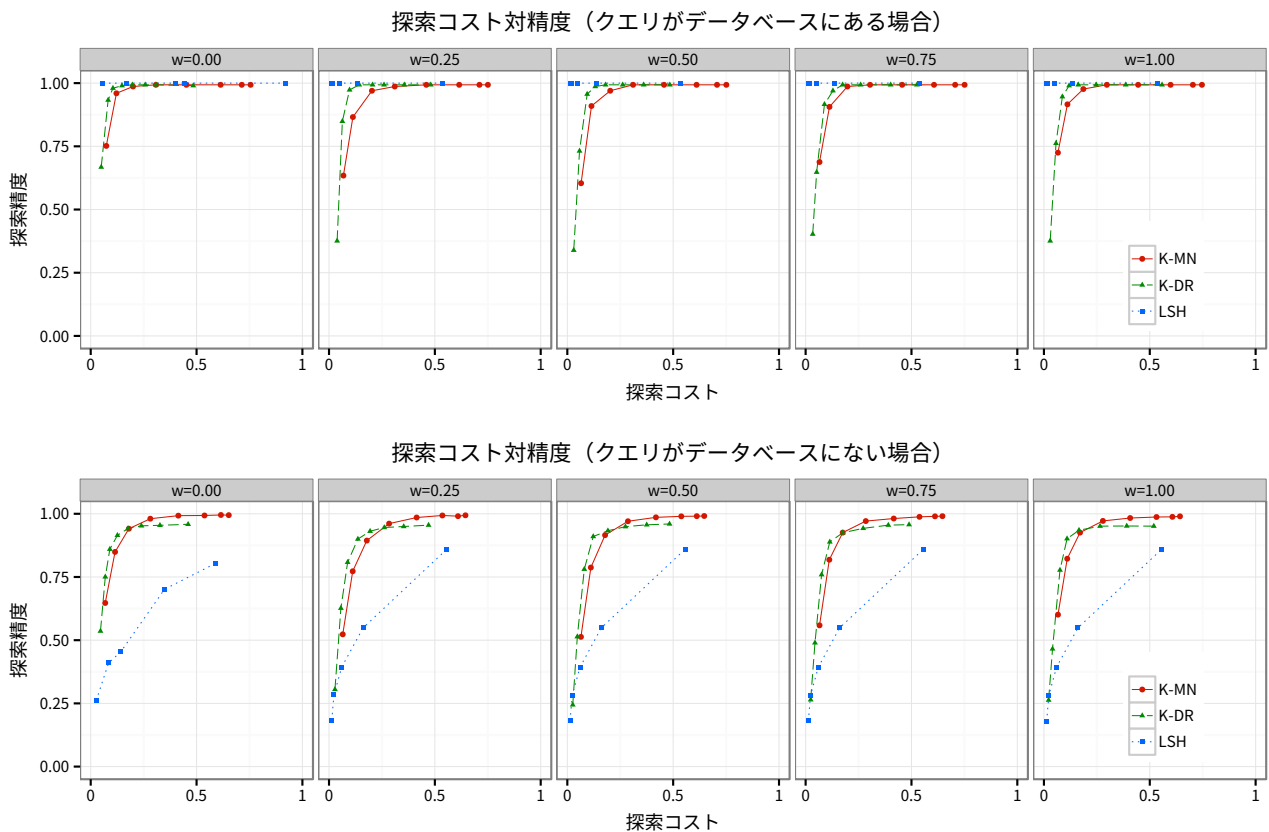


図 7 Holidays データセットに対する提案手法の探索コストと精度
(上: クエリがデータベースにある場合/下: クエリがデータベースにない場合)

用いた場合は樹木の葉のパターンが現れている画像が上位を占めている。一方で HSV カラーヒストグラムのみを用いた場合は、ほぼ暗闇の画像が上位を占めている。これらに対して、夜景中に電飾付き樹木があるというクエリ画像と同じ特徴を持つ画像は、全く同風景の画像を除けば、 $w = 0.5$ でのみ得られている。また 2 つ目の例では、Bag-of-Visual-Words を用いた場合はクエリ画像で花にピントがっていないためか、クエリ画像の特徴のうち地面のパターンにのみ反応した結果が上位を占めている。一方で HSV カラーヒストグラムのみを用いた場合はクエリ画像の花の色にのみ反応した結果が上位を占めている。これらに対し $w = 0.5$ で探索した結果は、それらの両方をバランスよく含む結果が得られている。

4.2 保証領域内の頂点の貪欲探索可能性

3 節末尾で述べた通り、 K -MN グラフにおいては補題 1 は必ずしも成立せず、保証領域 S_x 内の頂点から x まだが貪欲探索可能であるとは限らない。しかし、どの程度の確率で貪欲探索可能でなくなるかを実際に上記のデータベースに対して構築したグラフで検証したところ、99.98% の割合で貪欲探索可能であった。よって、保証領域内の頂点の貪欲探索可能性が厳密に保証されないことは、実用上問題が無いと考える。

4.3 探索コスト及び探索精度の検証

本節では、両データセットに対して次数パラメータ K を段階的に変化させながら探索精度とコストを検証した。また比較のため、単一非類似度に対する索引である K -DR グラフと Charikar らによる乱数超平面を用いた LSH[12] に対しても同様の実験を行った。ただしいずれも複数類似度に対応した索引ではないため、重視パラメータ w が与えられてから索引の作成を行っていることに注意されたい (LSH の場合は、 w が与えられてからその重みに従って特徴量を統合しハッシュ化している)。

なお、ここで言う探索コストとは全データ点数に対する探索の際に評価したデータ点数の割合であり、グラフ構築にかかるコストは含まれていない。評価したデータ点数とは、グラフ索引手法の場合は探索パスおよびその隣接頂点の総数であり、LSH の場合はハッシュ関数の次元数とハッシュテーブルから抽出された候補点数の和である。また探索精度とは、各クエリに対して探索結果が厳密探索の結果と一致した割合である。

また、実験に際しては 5 分割クロスバリデーションを行った。すなわち、データセットを 5 分割し、4/5 の訓練用データ集合の画像をデータベースに登録して索引を構築した。クエリ画像集合としては、データベースに登録されているデータから選ぶ設定と、データベースに登録されて

いない残り 1/5 のテスト用データ集合を用いる設定の両者を検証している。

実験結果は図 7 の通り。クエリ画像が既にデータベースに登録されている設定では、索引構築時にクエリに対する知識が活用できることになり、問題としてはより簡単なものとなっている。LSH はクエリがデータベースに登録されていれば必ずそのクエリを含むバケットを抽出できるため、ハッシュの次元数によらず探索精度は 100% になる。これに対し K -MN グラフでは、 $K = N - 1$ とすれば保証領域がすべての点を含むので、実質 100% の探索精度を達成できる (4.2 節参照)。また K -DR グラフの場合は、厳密に 100% である。しかしこれらのグラフ索引の性能は、より低い K に対しては LSH に劣る。

いっぽうクエリ画像がデータベースに登録されていないより現実に即した設定では、グラフ索引は LSH より高い性能を記録した。このうち K -MN だけは重視パラメータ w を使わずに構築された索引であるが、 w の各値に対して、 w が与えられてから構築された K -DR グラフと同等精度の性能を達成している。そのため、本稿で扱ったような複数の非類似度に対して柔軟に探索を行いたいという問題設定の場合、提案手法が有効であると言える。ただし辺の削減における基準が厳密でないため、まだ探索に冗長な辺が多く、高い精度を達成するには探索コストがやや大きくなりがちである。

5. おわりに

本稿では、複数の非類似度のある最近傍探索問題に対して探索時に非類似度間の重みを調整できる索引として K -MN グラフを提案した。 K -MN グラフは各非類似度における K 最近傍グラフを組み合わせた上で、探索に不要な辺を削減し探索効率を向上させたグラフ索引である。提案手法では索引の構築時には重視パラメータ w は指定しないにもかかわらず、あらゆる $w \in [0, 1]$ に対して一貫した性能を達成した。この性能は、従来の単一非類似度のみに対応する手法で重視パラメータを指定してから構築した索引に匹敵する性能であり、複数非類似度を取り扱う問題に対しても従来と同等程度の性能で最近傍探索が可能となる。

今後の課題として、提案手法を 3 つ以上の非類似度を扱えるよう拡張することが挙げられる。また利用者が自分の望む重視パラメータ w を数値として知ることが容易ではないため、これを利用者に入力させるのではなく、検索行動からリアルタイムに学習・調整できるような仕組みを開発することも課題である。

参考文献

- [1] Indyk, P. and Motwani, R.: Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, *Proceedings of the 30th annual ACM Symposium on Theory of Computing*, ACM, pp. 604–613 (1998).
- [2] Bentley, J. L.: Multidimensional Binary Search Trees Used for Associative Searching, *Communications of the ACM*, Vol. 18, No. 9, pp. 509–517 (1975).
- [3] White, D. A. and Jain, R.: Similarity Indexing with the SS-tree, *Proceedings of the 12th International Conference on Data Engineering*, IEEE, pp. 516–523 (1996).
- [4] Guttman, A.: R-trees: A Dynamic Index Structure for Spatial Searching, *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, Vol. 14, No. 2, pp. 47–57 (1984).
- [5] Yianilos, P. N.: Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces, *Proceedings of the 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, Vol. 93, No. 194, pp. 311–321 (1993).
- [6] Navarro, G.: Searching in metric spaces by spatial approximation, *The VLDB Journal*, Vol. 11, No. 1, pp. 28–46 (2002).
- [7] Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N.: *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley & Sons, Inc., 2nd edition (2000).
- [8] Sakagaito, J. and Wada, T.: Nearest First Traversing Graph for Simultaneous Object Tracking and Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1–7 (2007).
- [9] Sebastian, T. B. and Kimia, B. B.: Metric-Based Shape Retrieval in Large Databases, *Proceedings of the 16th International Conference on Pattern Recognition*, Vol. 3, IEEE, pp. 291–296 (2002).
- [10] Aoyama, K., Saito, K., Sawada, H. and Ueda, N.: Fast Approximate Similarity Search Based on Degree-Reduced Neighborhood Graphs, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1055–1063 (2011).
- [11] Datar, M., Immorlica, N., Indyk, P. and Mirrokni, V. S.: Locality-sensitive Hashing Scheme Based on p -stable Distributions, *Proceedings of the 20th Annual Symposium on Computational Geometry*, ACM, pp. 253–262 (2004).
- [12] Charikar, M. S.: Similarity estimation techniques from rounding algorithms, *Proceedings of the 34th Annual Symposium on Theory of Computing*, ACM, pp. 380–388 (2002).
- [13] Nister, D. and Stewenius, H.: Scalable Recognition with a Vocabulary Tree, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, IEEE, pp. 2161–2168 (2006).
- [14] Liu, C. and Wechsler, H.: A Shape-and Texture-Based Enhanced Fisher Classifier for Face Recognition, *IEEE transactions on Image Processing*, Vol. 10, No. 4, pp. 598–608 (2001).
- [15] Zhang, S., Yang, M., Cour, T., Yu, K. and Metaxas, D. N.: Query Specific Rank Fusion for Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 4, pp. 803–815 (2015).
- [16] Jégou, H., Douze, M. and Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search, *Proceedings of the 10th European Conference on Computer Vision: Part I* (Forsyth, D., Torr, P. and Zisserman, A., eds.), Lecture Notes in Computer Science, Vol. 5302, Springer, Springer, pp. 304–317 (2008).

クエリ画像



検索結果

$w = 0.00$ (Bag-of-Visual-Wordsのみ)

$d = 0.062$ $d = 0.156$ $d = 0.194$ $d = 0.195$ $d = 0.231$



$w = 0.50$

$d = 0.128$ $d = 0.256$ $d = 0.256$ $d = 0.342$ $d = 0.347$



$w = 1.00$ (HSVカラーヒストグラムのみ)

$d = 0.120$ $d = 0.121$ $d = 0.121$ $d = 0.121$ $d = 0.122$



クエリ画像



検索結果

$w = 0.00$ (Bag-of-Visual-Wordsのみ)

$d = 0.087$ $d = 0.094$ $d = 0.094$ $d = 0.097$ $d = 0.097$



$w = 0.50$

$d = 0.433$ $d = 0.459$ $d = 0.486$ $d = 0.499$ $d = 0.502$



$w = 1.00$ (HSVカラーヒストグラムのみ)

$d = 0.436$ $d = 0.533$ $d = 0.619$ $d = 0.649$ $d = 0.802$



図 8 w を 1.0,0.5,0.0 としたときの Holidays データセットに対する厳密探索解の例
それぞれ左段がクエリ画像，右段が検索結果を表す．検索結果は左から非類似度が小さい順に並んでいる．