

縮小構造 Sketch の焼きなまし法を用いた最適化における 近傍探索半径の選択に関する研究

小南 茜^{1,a)} 松永 ありす¹ 樋口 直哉¹ 今村 安伸¹ 篠原 武¹

概要：本研究では、多次元データの近似検索に用いる縮小構造 Sketch について検証する。Sketch は基礎分割関数を用いて作成され、実空間での類似性のある程度保持するようにオブジェクトをバイナリ文字列で表したものである。検索性能を向上させるために Sketch の最適化を試みる。異なるオブジェクトが同じバイナリ文字列になることを衝突といい、衝突が少ない程精度は高くなる。そこで、焼きなまし法の一種 AIR(Annealing by Increasing Resampling) を適用した。従来の焼きなまし法の温度にあたる部分をサンプルサイズとして調整し、最適化を行う。量子化球面分割 QBP により 2 値量子化して作られたピボットを用いて、近傍探索半径並びに 2 値のバランスを考慮したピボットの取り替え法の二つの探索手法を提案する。

キーワード：近似検索, 2 値量子化, 球面分割, 焼きなまし法

Study on Selection of Neighbor Search Radius in Optimization Using Annealing Method of Reduced Structure Sketch

AKANE KOMINAMI^{1,a)} ALICE MATSUNAGA¹ NAOYA HIGUCHI¹ YASUNOBU IMAMURA¹
TAKESHI SHINOHARA¹

Abstract: In this research, we examine the reduced structure Sketch used for approximate retrieval of multidimensional data. Sketch is created by using basic division function, and expresses objects in binary character string so as to hold similarity in real space to some extent, and attempts to optimize Sketch in order to improve search performance. The fact that different objects become the same binary character string is called a collision, and accuracy becomes higher as fewer collisions occur. Therefore, a kind of annealing method AIR was applied. The part corresponding to the temperature of the conventional annealing method is used as a sample and optimization is performed. We propose two search methods of pivot replacement with consideration of neighborhood search radius and binary balance by using pivot made by binary quantization with quantized spherical split QBP.

Keywords: approximate similarity search of multimedia data, ball partitioning functions, binary quantization, simulated annealing

1. はじめに

近年、計算機の演算能力及び記憶容量の飛躍的な向上により、我々は大量の音楽や動画などのマルチメディアデータを扱うことができるようになった。それに伴い、我々の日常生活では、それらの大量なデータに対して検索を行う機会が増加し、高速に検索することが重要になってきた。

音楽や動画といったマルチメディアデータを対象とした検索では、音質や画質の悪化が影響して、人間の感覚による判断では同一に思われる内容であっても計算機にとっては全く別のデータとして識別されるケースが非常に多い。そのため、それらのデータの高速な検索を実現するために、近似検索 [1], [2], [3] を用いる。近似検索とは蓄積されたデータの中から質問点と近似するオブジェクトを探索する検索方法である。

近似検索の一般的な検索手法は、R-tree[4] などの索引構

¹ 九州工業大学 情報工学部 知能情報工学科

^{a)} o231027a@mail.kyutech.jp

造を用いるものである。それらの空間索引は、空間の次元が大きくなると次元の呪いと呼ばれる現象により、検索効率が悪化してしまう。そのため、高次元のデータを用いて索引付けを行う場合、次元縮小を用いて、オブジェクトを低次元空間に射影することで、次元の呪いを緩和する。しかし、空間索引を用いた検索では、データベース内に近いデータが存在するような質問に対しては非常に高速に検索できる反面、近いデータが存在しない場合には検索効率が悪くなってしまふ。

そこで、近年考案された、高速かつ一定時間で検索を行うことが可能である Sketch[5], [6], [7] を用いた検索手法がある。Sketch は球面分割 (BP) [8] 等の基礎分割関数を用いることで、オブジェクトの類似度がある程度保持したまま、オブジェクトをバイナリ文字列で表現する。Sketch は、元のデータを小さく圧縮表現し、Sketch 間の距離はハミング距離で計算することで高速計算ができるため、高速な検索が実現できる。Sketch を用いた検索では、Sketch がフィルタリングのような役割として機能している。Sketch のデータベースに対して質問データから作成した Sketch との距離を計算し、近い順に K 個を候補とする。それらに対応するデータを取り出して実距離を計算して質問の近傍解を得る。このとき、 K 個の候補の中に最近傍解が含まれる確率が検索精度となる。Sketch を用いた検索では、高い検索精度を実現するために、Sketch 作成に用いる基礎分割関数のためのビットの選択が重要である。また、本来異なるオブジェクトが Sketch 上において同じバイナリ文字列になることを衝突と呼び、データベース全体で見たときに Sketch が衝突する確率を衝突確率とする。一般的に衝突確率が低い程、検索精度は高くなる。

本研究では、Sketch の最適化における近傍探索半径を変更することにより、Sketch の衝突確率の減少を目指す。2 章では Sketch とその検索手法について紹介し、3 章では焼きなまし法を用いた最適化手法について紹介する。4 章では今回提案した手法について紹介し、5 章ではその手法を用いて実際にデータベース内を検索し衝突確率と検索精度について検証し、6 章でまとめと考察を行う。

2. Sketch と近似検索

2.1 近似検索システム

近似検索システムとは、質問点に近似するデータをデータベースから取り出すシステムのことである。データ間に非近似度（距離）を定義し、質問点からの距離を順番でオブジェクトを取り出すことにより、近似検索を実現することができる。

近似検索システムにおいて、データベースが対象とする特徴空間全体を $\mathcal{U} = \mathbb{R}^n$ とする。ここで $\mathbb{R}$ は実数全体、 n は特徴の次元数を表す。任意の 2 点のオブジェクト間の非近似度の指標を示す距離関数を $d: \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}^+$ とし、

$\mathcal{D} = (\mathcal{U}, d)$ を距離空間とする。

近似検索システムでは、距離関数 d は距離の公理と呼ばれる次の条件を満たすものと仮定する。ただし同一性については、場合によっては満たさないこともあり、これを擬距離と呼ぶ。

- (1) $d(X, Y) \geq 0$ (非負性)
- (2) $d(X, Y) = d(Y, X)$ (対称性)
- (3) $d(X, Y) \leq d(X, Z) + d(Z, Y)$ (三角不等式)
- (4) $d(X, Y) = 0 \Leftrightarrow X = Y$ (同一性)

ここで、 $X, Y, Z \in \mathcal{D}$ である。上の条件の中で最も重要なものは、三角不等式である。

近似検索には、主に範囲質問および近傍質問の 2 種類の方法が用いられている。質問点 q と半径 $r \in \mathbb{R}^+$ を質問のパラメータとする範囲質問 $Range(\mathcal{D}, q, r)$ は、 q から距離 r 以内のオブジェクトを取得する質問である。すなわち、

$$Range(\mathcal{D}, q, r) = \{o \in S | d(o, q) \leq r\}$$

である。ここで、 S はデータベース内のオブジェクトの集合である。

また、質問点 q から距離が小さいものから k 個のオブジェクトを取得する質問 $NN(\mathcal{D}, q, k)$ を k 近傍質問という。さらに、上記の二つの質問方法を組み合わせた範囲限定 k 近傍質問 $k-NNRange(\mathcal{D}, q, r, k)$ がある。これは、質問点 q から距離が r 以内のものの中で、距離が近い順番にオブジェクトを k 個取得する。つまり、

$$k-NNRange(\mathcal{D}, q, r, k) = NN(\mathcal{D}, q, k) \cap Range(\mathcal{D}, q, r)$$

である。範囲限定最近傍質問は、一定の距離以上の解が必要であるときに用いると効果的である。

本論文で用いる距離計測について説明する。特徴空間内の任意のオブジェクトを x とする。 x の特徴は n 組の実数 $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ で表される。

以下の距離計測関数は距離の公理を満たす。

$$L_1 \text{ 距離} : D(x, y) = \sum_{i=1}^n |x_{(i)} - y_{(i)}|$$

$$L_2 \text{ 距離} : D(x, y) = \sqrt{\sum_{i=1}^n (x_{(i)} - y_{(i)})^2}$$

$$L_\infty \text{ 距離} : D(x, y) = \max_{i=1}^n |x_{(i)} - y_{(i)}|$$

2.2 Sketch

Sketch とは、実空間での類似性がある程度保持するように基礎分割関数を用いてオブジェクトをバイナリ文字列で表現したものである。オブジェクト x の Sketch を $\sigma(x)$ 、その第 i ビットを $\sigma_i(x)$ で表す。Sketch 間の距離は、従来、異なるビット数で表わされるハミング距離が用いられてきた。Sketch はバイナリ文字列で表現されているため、ハミ

ング距離はビット演算で高速に計算できる．本研究では，Sketch 間の距離は，距離下限を用いた手法 [9], [10] を使用する．

2.3 検索法

Sketch を用いた k 近傍検索は 2 段階で行う．

- (1) 質問データから作成した Sketch とデータベース内に登録されているデータの Sketch との距離を計算し，距離の小さな候補データを取り出す．
- (2) 取り出した候補データと質問データとの実空間上の距離を計算して解を求める．

最初の段階における Sketch データベースに対する検索では，全探索を用いて K 近傍質問 ($K \geq k \geq 1$) を行う．通常，全探索は索引構造を用いた場合と比較して，非常に検索に時間がかかるが，Sketch における全検索では，Sketch は元のデータに比べて小さく圧縮されており，距離を高速に計算することができるため，非常に高速に解を得ることが可能である．次に，得られた解を用いて，実際の距離に基づいた解を生成する [11]．Sketch は，実空間上の類似性を完全に保持できるわけではないため，Sketch を用いた検索では精度が悪化する．したがって，より良い類似性を保持した Sketch を作成するため基礎分割関数におけるピボットの選択が重要である．

2.4 基礎分割関数

本研究では，最適化手法として量子化球面分割法 (以下，QBP)[9] を用いる．QBP はランダムに選んだオブジェクトをデータの中央値によって 2 値量子化した点を中心点とし，中心点と中央値座標との距離を半径にする手法である．オブジェクト集合 $S \in \mathcal{U}$ と中心点 ($P \in \mathcal{U}$)，中心点と中央値座標との距離である半径 R が与えられたとすると， S は以下の様に部分空間 S_1, S_0 に分割される．

$$S_1 = \{o \in S | d(p_0, o) \leq r\}$$

$$S_0 = \{o \in S | d(p_0, o) > r\}$$

図 1 は，中心点 P' により，集合 S は， $S_1 = \{A, B, C\}$ と $S_0 = \{D, E, F\}$ に分割した例である．

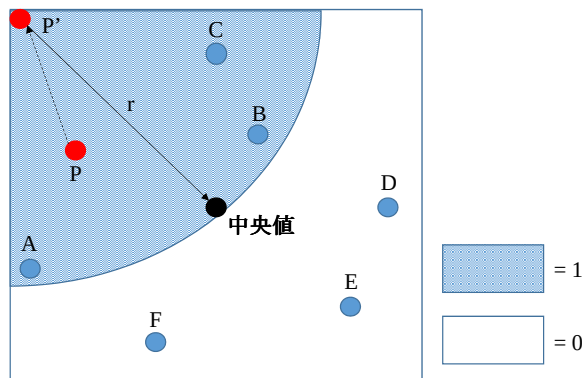


図 1 QBP

QBP において，Sketch の各ビット $\sigma_i(x)$ は，中心点 P と，半径 r を用いて，

$$\sigma_i(x) = \begin{cases} 1, & \text{if } d(p_i, x) \leq r \\ 0, & \text{if } d(p_i, x) > r \end{cases} \quad (i = 1, 2, \dots, t)$$

となる．

3. 焼きなまし法

物質を融解状態になるまで加熱し，徐々に冷却する操作を焼きなまし [12][13] という．焼きなましにより，エネルギーが最も少ない状態に分子が配列し，結晶構造を形成させることができる．この物理プロセスを計算機上で再現した技術が焼きなまし法と呼ばれている．

Sketch による近似検索においては，検索精度向上のために，ピボットの選択における最適化が重要である．ピボットの評価には最小衝突法を用いている．本来異なるオブジェクトが Sketch 上において同じバイナリ文字列になることを衝突と呼び，この衝突が少ないほど良いピボットと見なす評価手法である．一般的に Sketch の衝突が少ないほど検索精度は高くなる．

そこで衝突を減らすために本研究では，少数再サンプリングによる焼きなまし法 (以下，AIR)[9], [14] を用いる．従来の焼きなまし法では，初めに高い温度を与えランダムウォークさせ，温度をゆっくりと下げていくことで局所探索を行うようにする．AIR では，温度の代わりにサンプルサイズを変化させることで焼きなましを行う．高温時に少ないサンプルを用いることでランダムウォークを実現し，低温時に多くのサンプルを利用することで局所探索を実現している．ランダムウォークを行うことで，より良い中心点を検索し検索効率の上昇を図ることができる．ある程度試行したらサンプルの取り替えを行うが，そのタイミングはパラメータにより調整が可能である．この手法を用いると高温時に少ないサンプルのみを用いるため，従来の焼きなまし法よりも高速に最適化を行うことができる．

AIR のサンプルサイズの変化のスケジュール n は，進捗率を p ，初期サンプルサイズを n_0 とし， k_1 と k_2 の 2 変数により以下の式で調整を行う．このときの最大サンプルサイズを S_{max} とする．

$$n = \frac{n_0}{((1-p)^{k_1} \times 2^{-pk_2})^2} \quad (n \leq S_{max}) \quad (1)$$

(1) 式を用いた AIR によるサンプルサイズの変化の様子を図 2 に示す．

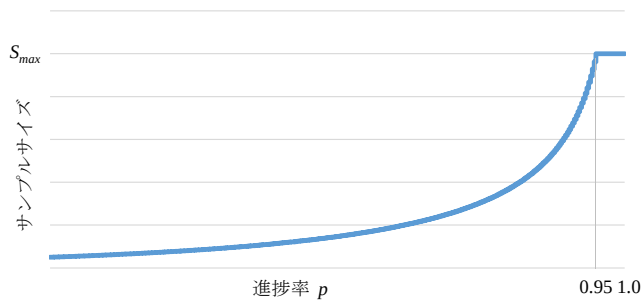


図2 AIR スケジュール

図2を見ると、初めは小さなサンプルサイズでランダムウォークさせ、終わりに急激にサンプルサイズを上昇させていることが確認できる。さらに、急激にサンプルサイズを上昇させたあとに、最大のサンプルサイズで局所探索を行っていることが確認できる。全体の試行回数を進捗率 p とし、残りの局所探索部分にかかる割合は $1-p$ となる。この進捗率 p を定めることで、(1) 式の k_1 と k_2 の2変数により、曲線部分の温度の上がり具合を調整することができる。

4. 提案手法

最適化に AIR を適用する中で、近傍に関する手法を提案する。まず AIR に関して、評価が同じでも遷移するようにし、より多くのランダムウォークを行うようにした。また、サンプルを取り替えると同時に置き換え対象となるピボットも選択する。このことにより、衝突を計算するうえでの効率化を図ることが可能である。

4.1 n 次元置き換え法 n -Exchange

特徴データの次元数を m とすると QBP によって揃えられた各ピボットの次元数は m 個となる。また、座標値はすべて空間の最大値または最小値である。ここでは、最適化における近傍探索半径の選択による探索の進み具合の変化を調査するために、 m 次元のうちの n 次元を取り替える n 次元置き換え法 (以下、 n -Exchange) という手法を提案する。このときの置き換える次元数 n を近傍探索半径とする。以下に手順を紹介する。

- (1) QBP により選ばれたピボットから差し替え元となるピボットを1つ選択する。
- (2) 選択したピボットの中から差し替え元となる次元を n 個選択する。
- (3) データベースからランダムにオブジェクトを1個取り出して量子化を行い、選択した n 個の次元の座標値をピボットにのものと置き換えながら Sketch を最適化する。
- (4) 評価が悪くならなければ遷移する。

データベースより新たに取得したオブジェクトの座標値と置き換えながら最適化を進めることで、衝突確率の減少

並びに検索精度の向上を期待した。

4.2 バランス置き換え法 Swap

QBP での基礎分割は、2値量子化した値の割合がおおよそ半分になっている点に着眼した近傍を提案する。このバランスを崩さないように、同一ピボットの2つの軸の値をお互い入れ替えるバランス置き換え法 (以下、Swap) を提案する。以下に手順を紹介する。

- (1) QBP により選ばれたピボットから差し替え元となるピボットを1つ選択する。
- (2) 選択したピボットの中から差し替え元となる次元と、今選んだ次元とは座標値が異なる次元、計2つの次元を選び、お互いの座標値を交換する。
- (3) 評価が悪くならなければ遷移する。

QBP で得られた、2値量子化した値のバランスを保持したまま最適化を進め、衝突確率の減少並びに検索精度の向上を期待した。

5. 実験

実験データとして、約2,900本の動画から64次元に特徴抽出した約700万件の画像データを用いる。質問データとしては、近質問、準近質問、遠質問、各100件の計300件のデータを用いる。このデータに対して、基礎分割関数として QBP を用いて 32bit の Sketch を作成し、提案手法である n -Exchange と Swap を適用する。Sketch によるフィルタリングにおける解候補数 $K=7,000$ とする。このときの計算時間は最適化にかかる時間で、検索精度と衝突確率を比較に用いる。また、最大サンプルサイズを10万とし、そのサンプルを用いて再サンプリングを行う。比較にはデータベースより取得するサンプルデータを変えて各10回試行したものの平均を用いる。

AIR の温度スケジュール n は、初期サンプルサイズを500として、(1) 式より $K_1 = 0.66$, $K_2 = 1.0$ とし、総試行回数の95%で最大のサンプルサイズに到達し、残りの5%では最大サンプルサイズで局所探索を行うように調整する。

5.1 近傍探索半径 n の変更

n -Exchange を用いて近傍探索半径を変化させながら検証を行った。最適化の試行回数は50万回とし、実験結果は、表1に示す。

表1 n -Exchange の適用 (試行回数: 50万回)

n	衝突確率	時間	検索精度
64	2.31×10^{-6}	74.4 秒	97.1 %
32	2.14×10^{-6}	69.0 秒	96.9 %
16	2.03×10^{-6}	66.5 秒	97.1 %
8	2.02×10^{-6}	65.4 秒	97.0 %
4	2.00×10^{-6}	64.9 秒	96.8 %
2	2.02×10^{-6}	65.0 秒	96.7 %

表 1 より、近傍探索半径を小さくするにつれて 64 から 4 までは、衝突確率が減っていることが確認できた。しかし、近傍探索半径が 2 のときには衝突確率が上がってしまった。これより、近傍探索半径を大きくし過ぎると最適解を通り過ぎてしまい、小さくし過ぎると最適化にかかる試行回数が足りず、最適解までたどり着けていないと思われる。

5.2 バランス取り替え法

Swap を適用した検証を行った。実験結果を表 2 に示す。

表 2 Swap の適用

試行回数	衝突確率	時間	検索精度
10 万回	2.11×10^{-6}	12.4 秒	96.8 %
50 万回	2.01×10^{-6}	58.5 秒	96.7 %
100 万回	2.00×10^{-6}	114 秒	97.1 %

また、Swap と n -Exchange の効果を比較するために、同試行回数で n -Exchange も実施した。このときの、 n -Exchange の探索近傍半径は実験 5.1 で最も低い衝突確率を達成していた 4 を採用した。実験結果を表 3 に示す。

表 3 n -Exchange の適用 ($n=4$)

試行回数	衝突確率	時間	検索精度
10 万回	2.11×10^{-6}	13.9 秒	97.0 %
50 万回	2.00×10^{-6}	64.9 秒	96.8 %
100 万回	1.96×10^{-6}	128 秒	96.7 %

表 2 より、試行回数を増やすほど衝突確率が減っていることが確認できた。しかし、試行回数 50 万回と 100 万回とでは衝突確率は大幅に減っているわけではないため、最適化の限度に達していると考えられる。

表 2 と表 3 を比較すると、試行回数を大きくすると n -Exchange の方が衝突確率が減っていることがわかる。Swap のほうが数秒早い結果となっているのは、 n -Exchange が差し替える際、量子化にかかるコストの差が原因であると考えられる。

5.3 従来手法との比較

従来最適化手法として用いる QBP と、本研究の提案手法である n -Exchange と Swap の最適化の進み具合を比較する。3 つの手法とも同等の計算コストとなるように試行回数を調整して比較を行った。また、このときの、 n -Exchange の探索近傍半径は実験 5.1 で最も低い衝突確率を達成していた 4 を採用した。実験結果は、表 4 に示す。

表 4 より、衝突確率を減らすという意味では n -Exchange と Swap は大いに有効であることがわかる。

表 4 従来手法 QBP と n -Exchange, Swap の比較

最適化手法	衝突確率	時間	検索精度
QBP	2.74×10^{-6}	63.4 秒	97.2 %
n -Exchange	2.00×10^{-6}	64.9 秒	96.8 %
Swap	2.01×10^{-6}	58.5 秒	96.7 %

6. まとめと考察

近傍探索半径の変更実験においては、近傍探索半径が 4 のときに最も低い衝突確率を達成しており、最適な近傍探索半径は 4 であることがわかった。近傍探索半径を小さくすれば衝突確率は減ったが、小さ過ぎると最適解までたどり着くのに時間がかかってしまい、試行回数が足りないことが確認できる。よって、近傍探索において適切な歩幅の調節が必要であることが分かる。

また、Swap と n -Exchange で最適な近傍探索半径であった $n=4$ を比較すると、十分な試行回数を重ねることで Swap よりも n -Exchange のほうが衝突確率は減っていた。 $n=4$ のときの n -Exchange と Swap が似たような結果になったのは、 n -Exchange でデータベースより新たに取得したオブジェクトの座標値と取り替える際、取り替える前と取り換えた後で、本当の意味で取り替わっている軸の数が Swap とほぼ同じになっていることが考えられる。

最適化時間においては、量子化のコストから Swap のほうが速く検索できる。より速い時間で衝突確率を減らす、という点においては Swap は優れているが、検索精度の点ではどちらが優れているとはいえない。本研究で提案した手法の n -Exchange では、今回はデータベースからランダムに選択したオブジェクトの座標値を量子化したものの入れ替えを行ったが、このときのオブジェクトの選択方法を改良することでより効率の良い近傍探索ができると考える。

どの実験においても検索精度に大きな差はなく、衝突確率が低いほうが検索精度が高くなっていることは確認できなかったため、最適化を進め、衝突を減らすことで検索精度を上げるという点の限界が見えている。検索精度を上げるためには、衝突確率を減らすという指標ではなく、別の指標を用いて最適化する手法を検証する必要がある。そして、本研究では画像データでのみ検証を行っているため、音データやランダムデータなどの他のデータでの検証が必要である。

参考文献

- [1] 青木隆明, 岩崎瑤平, 篠原武: 棋譜データベースからの近似盤面検索に対する効果的な射影法に関する研究, 火の国情報シンポジウム, 2009.
- [2] 浦郷祐希, 田島圭, 青木隆明, 岩崎瑤平, 篠原武: 空間索引による近似画像の高速検索を用いた動画同定システムの実現, 火の国情報シンポジウム, 2009.
- [3] 岩崎瑤平, 篠原武, A.J.Muller: Sketch による大容量記憶データに対する高速な空間検索法に関する研究, 第 74 回 SIG-FPAI, 人口知能学会, 2009.
- [4] Guttman, A: R-trees: A dynamic index structure for spatial searching, Proc. ACM SIGMOD, International Conference on Management of Data, pp. 47-57, 1984.
- [5] Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures, In CIKM '04, pages 208-217, New York, NY, USA, 2004. ACM.
- [6] Lv, Q., Josephson, W., et.al: Efficient filtering with sketches in the ferret toolkit In MIR '06, pages 279-288, New York, NY, USA, 2006. ACM.
- [7] Muller, A.J., T. Shinohara, T.: Efficient Similarity Search by Reducing I/O with Compressed Sketches, International Workshop on Similarity Search and Applications, pp.30-38, 2009.
- [8] 樋口直哉 ほか, 2 値量子化に基づく球面分割を用いた類似検索のための Sketch, 九州支部連合大会, 2016
- [9] 樋口直哉, 今村安伸, 篠原武: 縮小構造 Sketch と質問点との距離下限値を利用した近似検索の高性能化に関する研究, 火の国シンポジウム, 2017.
- [10] Higuchi, N., Imamura, Y., et.al: Nearest neighbor search using sketches as quantized images of dimension reduction, In Proc. ICPRAM'18, pp. 356-363, 2018.
- [11] 岩崎瑤平, 篠原武: Sketch を用いた空間検索法の精度向上に関する研究, 第 76 回 SIG-FPAI, 人工知能学会, 2010.
- [12] Kirkpatrick, S., Gelett Jr. C. D., Vecchi, M. P. Optimization by Simulated Annealing. Science, 1983.
- [13] (社) 日本熱処理技術協会/日本金属熱処理工業会 編集, 熱処理技術入門, pp.150-156, 大河出版, 1974
- [14] Imamura, Y., Higuchi N., et.al: Pivot selection for dimension reduction using Annealing by Increasing Resampling, In Proc. LWDA'17, pp. 15-23, 2017.