DEIM Forum 2009 B6-3

類似度のランクに基づくクラスタリング手法と類似検索への応用

小久保侑紀† 星 守† 小早川倫広†† 山中 克久† 大森 匡†

† 電気通信大学大学院情報システム学研究科 〒 182–8585 東京都調布市調布ヶ丘 1–5–1
†† 沖縄工業高等専門学校 メディア情報工学科 〒 905–2192 沖縄県名護市辺野古 905
E-mail: †{kokubo,hoshi,yamanaka}@hol.is.uec.ac.jp, ††kobayaka@okinawa-ct.ac.jp

あらまし 大量のデータから有用な情報を解析するために様々なクラスタリング手法が提案されている. 従来手法の多くは, 距離の公理を満たす類似度に基づいてクラスタリングを行っている. しかし, 実際には距離の公理を満たさないものも多く存在し, 従来のクラスタリング手法が適用できないことも多い. 金城らはそのような距離の公理を満たさない類似度に対して適用できるクラスタリング手法とそれを用いた類似検索手法を提案した. その手法は, 類似度のランクがお互いに高いデータ同士にエッジを張ることにより, 各データをノードとするグラフを作成し, そのエッジ情報に基づきクラスタリングを行うものである. 本稿では金城らの手法のマージ処理を改良し, 高精度な類似検索を実現するクラスタの作成手法と検索アルゴリズムを新たに提案する. 領域画像データを用いたクラスタリング実験, 及び, 作成したクラスタを用いた検索実験を行い, 類似度計算回数, 及び再現率を求めた. その結果から本手法を用いることにより, クラスタ内のノイズを減少させ, 高速かつ高精度な検索を行えることが確認できた.

キーワード クラスタリング, k-MNN グラフ

A Rank-based Clustering Algorithm for Non-metric Space and its Application to Retrieval

Yuuki KOKUBO†, Mamoru HOSHI†, Michihiro KOBAYAKAWA††, Katsuhisa YAMANAKA†, and Tadashi OHMORI†

† The University of Electro-Communications, Graduate School of Information Systems, Chofugaoka 1–5–1, Chofu, Tokyo, 182–8585 Japan
†† Okinawa National College of Technology, Department of Media Information Engineering, Henoko 905, Nago, Okinawa 905–2192
E-mail: †{kokubo,hoshi,yamanaka}@hol.is.uec.ac.jp, ††kobayaka@okinawa-ct.ac.jp

Abstract Various clustering algorithms are proposed to extract useful information from a large amount of data. Kinjo et al. proposed a graph theoretical clustering algorithm for a non-metric similarity and its application to image retrieval. The key idea is that data in the same group are mutually in the top k rank with respect to the similarity. In this paper we improve the merging process and propose a new clustering algorithm. We apply our algorithms to the set of 34000 segmented images. Then we evaluate the number of computations of the similarity and recall ratio. The results showed that efficiency and accuracy of our algorithms.

Key words Clustering, k-MNN Graph

1. はじめに

大量のデータから有用な情報を解析するために様々なクラスタリング手法が提案されている [1] [2] [3] [4] [5]. 従来手法の多くは, 距離の公理を満たす類似度に基づいてクラスタリングを行っている. しかし, 類似度の中には距離の公理を満たさないものもある. そのような類似度に対して, データの構造, 分布, データ間の類似度の性質に則したクラスタリング手法も提案されている [3] [4].

金城ら [6] [7] は距離の公理を満たさない類似度に対して適用できるクラスタリング手法とクラスタを用いた類似検索を提案している. 彼らは領域画像データに対して, 画像同士の類似度として重み付け Jaccard 係数を用いて領域画像データのクラスタ作成実験を行っている. 本稿では金城らの用いた大規模なデー

タベースに対し, クラスタ作成実験を行い, クラスタ内の特性を分析する. その結果から, クラスタ内のデータ間の類似度の最大値に着目し, クラスタ内のノイズを少なくするマージ手法を提案する.

本稿の構成は次の通りである. 2 節では準備として, 金城ら [6] の提案しているクラスタリング手法 (以下では従来手法と呼ぶ) について述べる. また, 34000 枚の領域画像を用いてクラスタ作成実験を行い, クラスタを分析する. 3 節では, クラスタのマージ処理部分を改良した手法 (改良手法と呼ぶ) を提案し, クラスタ作成実験を行い, 従来手法と比較する. 4 節ではクエリとクラスタの代表点間の類似度を用いた検索アルゴリズムを提案し, 検索実験を行い, 類似度計算回数, 再現率を分析する. 5 節では類似度計算回数を更に削減させるための索引構造について述べる. 6 節では検索の特徴を調べるため, クエリ集合を変更し検索実験を行う. 7 節でまとめを述べる.

2. 金城らの手法

本節では準備として, 金城ら [6] の提案したクラスタリング手法について述べる. 彼らの手法は, 互いに似ているもの同士はランクが高いという考えに基づき, k-MNN グラフを定義し, そのエッジの情報に基づいてクラスタリングを行うものである.

2.1 k-MNN グラフ

次のように k-MNN グラフ (k-mutually Nearest Neighbor Graph) を定義する. N 個のデータからなる集合を $\mathcal{D} = \{d_1, \ldots, d_N\}$ とする. $\mathcal{D}$ の各要素 $d_i (i = 1, \ldots, N)$ に対してノード n_i を割り当てる. ノード n_i と n_j との類似度を $sim(n_i, n_j)$ と書くことにする. $sim(n_i, n_j)$ が $sim(n_i, n_1), sim(n_i, n_2), \ldots, sim(n_i, n_N)$ のうち X 番目に大きいならば, $simRank(n_i, n_j) = X$ と書き, n_i から n_j への $simRank$ は X であると言う. このとき, 定数 k と $\mathcal{D}$ に対する k-MNN グラフとは, ノード集合 $V = \{n_1, \ldots, n_N\}$ とエッジ集合 $E = \{(n_i, n_j) | simRank(n_i, n_j) < k$ かつ $simRank(n_j, n_i) < k\}$ からなるグラフである. k-MNN グラフを用いたクラスタリングの手順をアルゴリズム **Algorithm 1** に示す.

Algorithm 1 クラスタリング手順

手順 1(k-MNN グラフ作成) $\mathcal{D}$ 中の全てのデータ間で類似度を計算し, k-MNN グラフを作成する.

手順 2(極大クリーク抽出) k-MNN グラフ中の全ての極大クリークを抽出する. 得られた極大クリークを初期クラスタ集合とする.

手順 3(マージ処理) 初期クラスタ集合に対し, 式 (1) を用いて, マージ処理を行う (詳細は後述).

手順 3 で得られるクラスタ集合 $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ をクラスタリング結果とする. マージ処理については次節で説明する.

2.2 マージ処理

クラスタのマージ処理について説明する. k-MNN グラフの極大クリークを抽出しているため, 同じデータが複数のクラスタに属することがある. 金城らはこのことを利用し, 共通要素の多いクラスタ同士をマージさせている. 具体的には, クラスタ $\mathcal{C}_p, \mathcal{C}_q$ が以下の式

$$\frac{|\mathcal{C}_p \cap \mathcal{C}_q|}{min(|\mathcal{C}_p|, |\mathcal{C}_q|)} \geqq \varepsilon \tag{1}$$

を満たすとき, $\mathcal{C}_p$ と $\mathcal{C}_q$ をマージする. ここで, $|\mathcal{C}_p|, |\mathcal{C}_q|, |\mathcal{C}_p \cap \mathcal{C}_q|$ はそれぞれのクラスタに属する要素の個数, ε は閾値である. (1) 式を満たすクラスタの組が無くなるまでマージを繰り返し, 結果として得られるクラスタ集合をクラスタリング結果とする.

2.3 クラスタ作成実験

作成されるクラスタの特徴を分析するため, クラスタ作成実験を行った. 本実験では金城ら [6] の用いたデータセットを拡張した 34,000 枚の領域画像データを用いた.

[重み付き Jaccard 係数]

クラスタ作成実験で用いる類似度である重み付き Jaccard 係数について述べる. 構図による類似画像検索のための領域の重なり度合いを測定する尺度として, 山本ら [8] は, Jaccard 係数に基づく新しい類似度 (重み付き Jaccard 係数に基づく類似度)

$$sim(I_i, I_j) = \sum_{R_{ip} \in I_i} \sum_{R_{jq} \in I_j} \frac{|R_{ip} \cap R_{jq}|^2}{|R_{ip}| \times |R_{jq}|} \times \frac{|R_{ip} \cap R_{jq}|}{|R_{ip} \cup R_{jq}|} \tag{2}$$

を提案している. ここで, R_{ip}, R_{jq} は領域画像 I_i, I_j を構成する領域であり, $|R|$ は領域 R に属する画素数である. この重み付き Jaccard 係数に基づく類似度は, 距離の公理のうち三角不等式を満たしていない尺度である.

[クラスタ作成実験の結果]

クラスタ作成実験について述べる. k-MNN グラフ作成時のパラメータを $k = 5$ とし, ε の値を変化させたときのクラスタの個数 (クラスタ数), 1 つのクラスタに属する領域画像数 (クラスタサイズ) の平均, 分散 (サイズ平均, サイズ分散), 1 枚の領域画像のみからなるクラスタ (孤立点) の個数を表 1 に示す. ε の値を大きくするにつれて, クラスタ数は大きくなるが, 孤立点となる画像の数は変わらない. 極大クリークを抽出しているため, 孤立点となる画像が複数のクラスタに属することがないからである. 図 1 に $k = 5, \varepsilon = 0.5$ でクラスタを作成したときのクラスタサイズの相対頻度分布を示す. 横軸はクラスタサイズ, 縦軸は相対頻度である.

表 1 従来手法で作成されるクラスタ ($k = 5$)

ε	クラスタ数	サイズ平均	サイズ分散	孤立点
0.2	16187	2.17	8.14	11604
0.5	17157	2.10	5.01	11604

作成されたクラスタを見ると, 人間の目で見て, 明らかに他とは似ていない画像が混じっているものがある. 図 2 にそのようなクラスタの例を示す. このクラスタは, $k = 5$, $\varepsilon = 0.5$ としたときに得られたものであり, 9 つの領域画像を含んでいる. 画像 1 から画像 9 への $simRank$ を求めたところ 31403 となった. これは, 全体 (34000) の中でも非常に大きな値である. 複数のクラスタにおいて, 同様に $simRank$ が大きくなる画像の組が観測された. $simRank$ が大きい画像が含まれているクラスタ

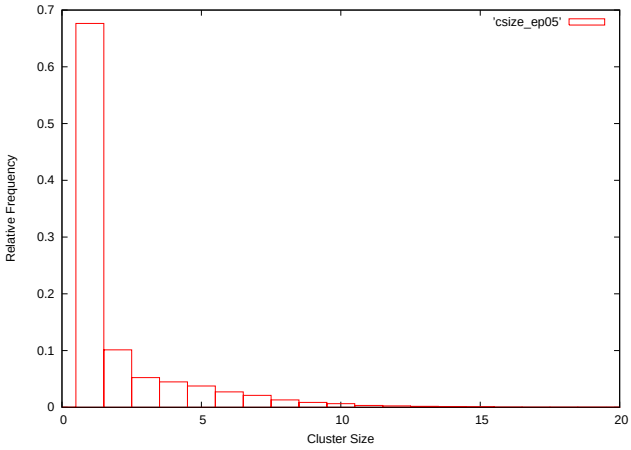


図 1　クラスタサイズの相対頻度分布 ($k = 5, \varepsilon = 0.5$)

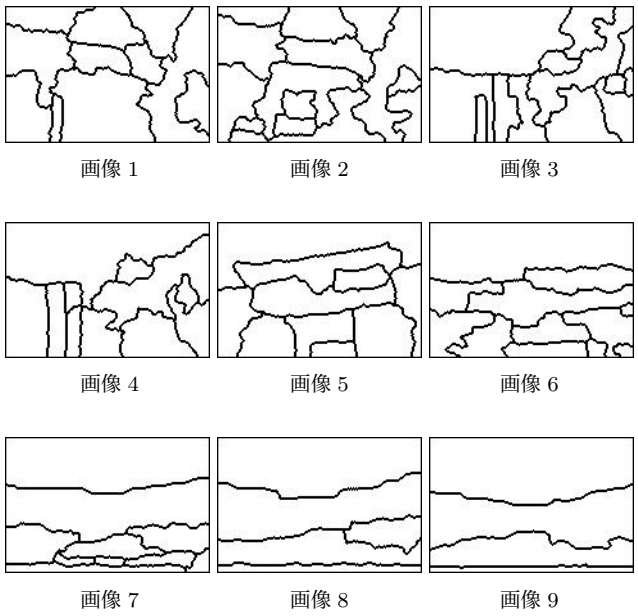


図 2　従来手法で作成される良くないクラスタの例

は人間の目で見たとき, 複数のグループが混在しているものや, クラスタ内で明らかに他と違っている “ノイズ” を含むものが多い. このようなクラスタは望ましくない. このようなクラスタが生成されないようにするための改良手法を次節で述べる.

3. 改 良 手 法

3.1 マージの改良

従来手法のマージ手法では, $simRank$ が大きい画像の組を含むクラスタを生成してしまうことがある. そこで, クラスタ内のデータに対する $simRank$ に着目し, マージ処理を行う手法を提案する. 改良手法には, 次に示す $simRank$ 最大値を用いる.

[$simRank$ 最大値]

各クラスタ $\mathcal{C}_i (i = 1, \ldots, M)$ に対し, $simRank$ 最大値 ($Max\_simRank$) を

$$Max\_simRank(\mathcal{C}_i) = \max_{x,y \in \mathcal{C}_i} simRank(x, y)$$

と定義する.

Algorithm 2 改良クラスタリング手順

手順 1(k-MNN グラフ作成) $\mathcal{D}$ 中の全てのデータ間で類似度を計算し, k-MNN グラフを作成する.

手順 2(極大クリーク抽出) k-MNN グラフ中の全ての極大クリークを抽出する. 得られた極大クリークを初期クラスタ集合とする.

手順 3(マージ処理) 初期クラスタ集合に対し, 式 (3) を用いてマージ処理を繰り返し行う.

表 2　改良手法で作成されるクラスタ ($k = 5$)

δ	クラスタ数	サイズ平均	サイズ分散	孤立点
50	18137	2.44	1.69	2568
100	15831	2.79	2.50	999
500	11394	3.93	7.65	39
1000	9322	4.84	13.14	4
1500	8097	5.58	19.20	0
2000	7247	6.24	23.99	0
2500	6667	6.78	30.69	0
3000	6168	7.33	35.10	0

2.2 節の式 (1) の代わりに, $\mathcal{C}_p$, $\mathcal{C}_q$ が以下の式を満たすとき, クラスタ $\mathcal{C}_p$ と $\mathcal{C}_q$ をマージする.

$$Max\_simRank(\mathcal{C}_p \cup \mathcal{C}_q) \leqq \delta \qquad (3)$$

ここで δ は閾値である.

(3) 式を満たすクラスタの組が複数存在する場合は, $simRank$ 最大値が小さいものからマージを行う. (3) 式を満たすクラスタの組が無くなるまでマージを繰り返し, 結果として得られるクラスタ集合をクラスタリング結果とする. **Algorithm 2** に改良手法の大まかな手続きを示す.

3.2 実 験 結 果

改良手法でのクラスタ作成実験を行った. $k = 5$ とし, δ の値を変化させたときのクラスタの個数 (クラスタ数), 1 つのクラスタに含まれる領域画像数の平均, 分散 (サイズ平均, サイズ分散), 1 枚の領域画像のみからなるクラスタ (孤立点) の個数を表 2 に示す. δ が大きくなるにつれて, クラスタ数が小さくなり, 1 つのクラスタに含まれる要素数が大きくなることが分かる. また, 孤立点の個数が減少している. 図 3 に $k = 5, \delta = 100$ でクラスタを作成したときのクラスタサイズの相対頻度分布を示す.

[$simRank$ 最大値の比較]

作成したクラスタから $simRank$ 最大値を算出したものを図 4 に示す. パラメータを $\varepsilon = 0.2$, $\varepsilon = 0.5$ として従来手法を用いた結果を図 4 中の ep02, ep05 にそれぞれ示す. また, $\delta = 100$ として改良手法を用いた結果を d100 に示す. 縦軸は割合, 右の数字は $simRank$ 最大値の値の範囲を示す.

従来手法では, $simRank$ 最大値が大きいクラスタが多数存在している. $\varepsilon = 0.2$ のクラスタでは 4 割以上のクラスタで $simRank$ 最大値が 1001 以上となっており, 10001 以上となるクラスタも複数存在している. 改良手法では $\delta = 100$ に抑えられており, 9 割程度のクラスタの $simRank$ 最大値は 50 以下となっている. 全体的に $simRank$ が高いクラスタ同士がマージされているのが分かる. ($\delta = 50$ のときは全てのクラスタで

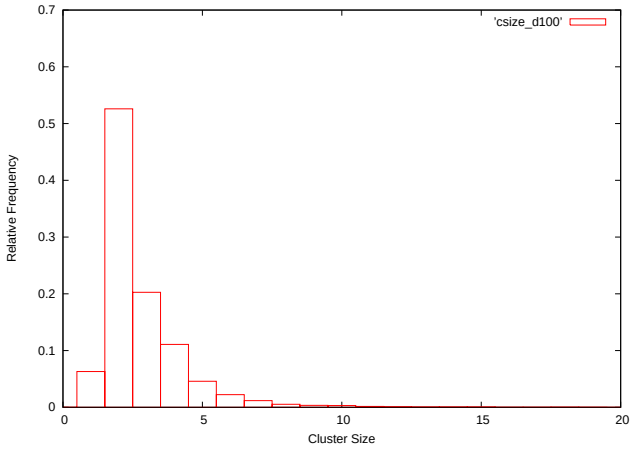


図 3　クラスタサイズの相対頻度分布 ($k = 5, \delta = 100$)

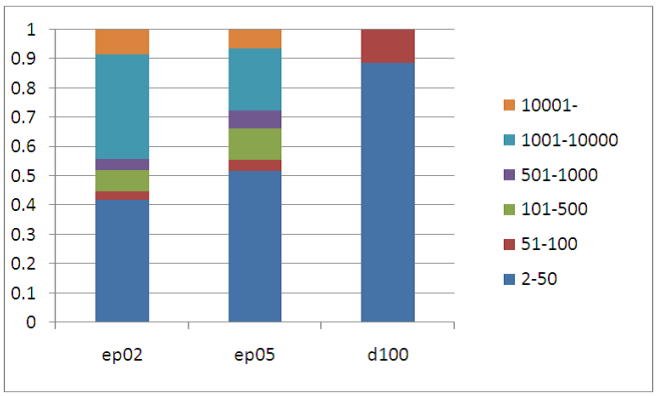


図 4　$simRank$ 最大値の比較

$simRank$ 最大値は 50 以下となる.)

[孤立点の比較]

従来手法では k-MNN グラフで孤立点となる画像はマージの閾値 ε を変えても孤立点のままである. 改良手法では k-MNN グラフで孤立点となる画像もマージされることが生じるようになり, δ が大きくなるにつれて孤立点の個数は減少する. また, $\delta = 50$ でのマージでも孤立点の個数は 2568 個と, 従来手法と比べ大幅に削減されている.

図 5 は, 改良手法 ($k = 5, \delta = 100$) で作成されるクラスタの例である. 従来手法では孤立点となった画像 1 がマージされている.

4. 検索実験

作成したクラスタを利用して, クエリに対し, 次節で述べる検索アルゴリズムを用いて, 類似度が上位 l 以内の画像を取り出す検索実験を行う.

4.1 検索アルゴリズム

クラスタ $\mathcal{C}_i$ 内の全ての点に対し, 類似度の和 $\sum_{x \in \mathcal{C}_i} sim(p_i, x)$ が最大となる点を $\mathcal{C}_i$ の代表点 p_i とする. 検索の手順を **Algorithm 3** に示す.

[再現率]

クエリに対しデータベース中の全ての要素とクエリとの類似度を計算し, 上位 l 以内の要素を求め, 得られた要素からなる集合を正解集合 $Truth$ とする. また, 検索によって得られる結果の集合を $Result$ とすると, 再現率は以下の式で表される.

$$\frac{|Result \cap Truth|}{|Truth|}$$

となる. ここで, $|Truth| = l$ である.

Algorithm 3 検索アルゴリズム

手順 1(クラスタの選択) 作成した各クラスタ $\mathcal{C}_i (i = 1, \ldots, M)$ に対し, クエリと代表点 p_i との類似度を計算し, 類似度の大きいものから上位 m 以内のクラスタを選択する. m を選択クラスタ数と呼ぶ.

手順 2(検索リストの作成) 手順 1 で選択した各クラスタ $\mathcal{C}_j$ について, $\mathcal{C}_j$ の要素を検索リスト A に追加する.

手順 3(類似画像の出力) A に含まれる全ての要素とクエリとの類似度を計算し, 上位 l 個を出力結果とする.

4.2 実　　験

データベース全体からランダムに 1000 枚の画像を選び, テストクエリとする. 各クエリ対して上記のアルゴリズムを適用し, 再現率を求め, 1000 枚のクエリでの平均を算出した.

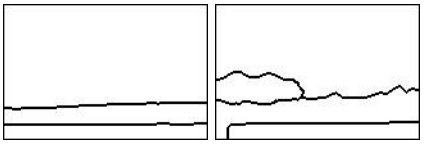
[選択クラスタ数と再現率の比較]

選択クラスタ数 m の範囲を 1 から 300 とし, 全クエリに対する再現率の平均を求めた. 出力する画像数を $l = 5$, $l = 20$ とし, 検索を行った結果を図 6, 7 にそれぞれ示す. 横軸は選択クラスタ数 m, 縦軸は再現率の平均である. 赤, 緑, 青, 紫の点はそれぞれ, 改良手法 ($k = 5$, $\delta = 50$), 改良手法 ($k = 5$, $\delta = 100$), 従来手法 ($k = 5$, $\varepsilon = 0.5$), 従来手法 ($k = 5$, $\varepsilon = 0.2$) の結果を表している.

l を大きくすると全てのクラスタにおいて, 再現率は下がる. 改良手法と従来手法を比較すると, 改良手法の方が再現率は高くなっている. すなわち, 改良手法の方が少ない選択クラスタ数で, 高精度な検索を実現している.

画像 1

画像 2　　画像 3



画像 4　　画像 5

図 5　従来手法で孤立点となり, 改良手法でクラスタ化される例

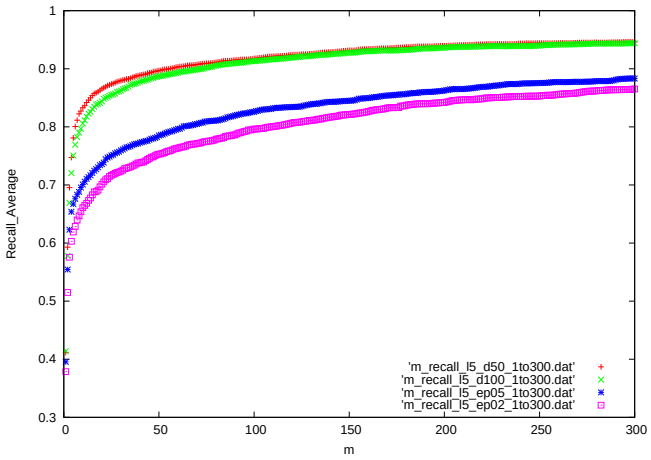


図 6　選択クラスタ数 m と再現率の平均 ($l = 5$)

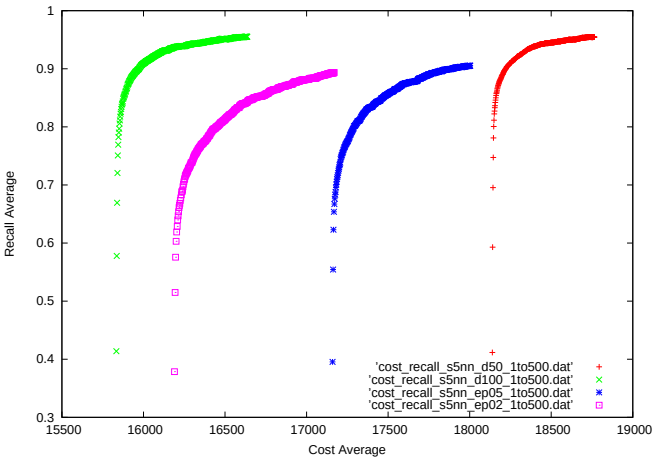


図 8　類似度計算回数の平均と再現率の平均 ($l = 5$)

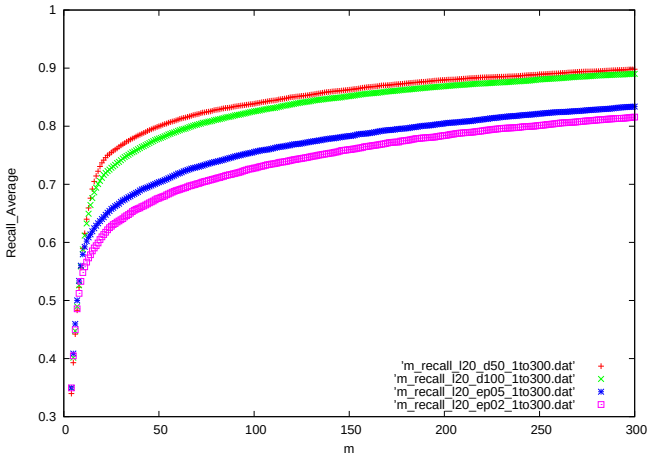


図 7　選択クラスタ数 m と再現率の平均 ($l = 20$)

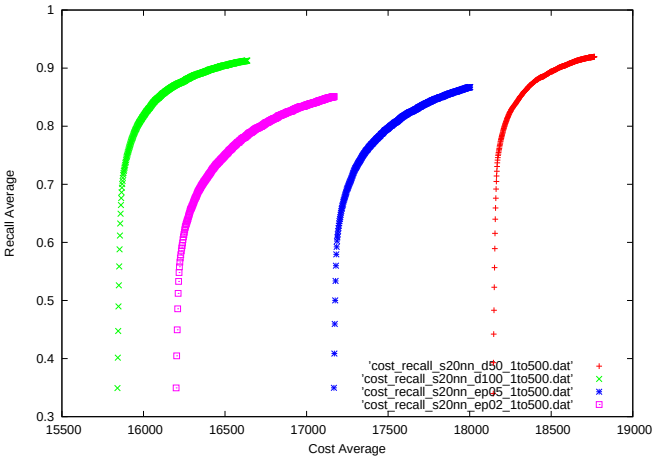


図 9　類似度計算回数の平均と再現率の平均 ($l = 20$)

[類似度計算回数と再現率の比較]

1 つのクエリに対して実行される類似度計算の回数を考える．クエリとの類似度計算は, 各代表点とクエリとの類似度計算に M 回を要し, さらに検索リスト A の中を全検索するので $|A| - m$ 回必要となる (代表点との計算は既に行っているため m を引く)．よって, クエリとの類似度計算回数は $M + |A| - m$(回) となる．

出力する画像数を $l = 5$, $l = 20$, 選択クラスタ数 m の範囲を 1 から 500 とし, 全クエリに対する類似度計算回数の平均, 及び再現率の平均を求めたものを図 8, 9 に示す．横軸は類似度計算回数の平均, 縦軸は再現率の平均を表す．グラフにおいて, 左上に布置するクラスタの方が, より少ない比較回数で, より高精度の検索を行えることを示している．図 8 より改良手法 ($k = 5$, $\delta = 100$) は従来手法に比べ, 類似度計算回数が抑えられれ, 高精度な検索が実現できていることがわかる．しかし, $\delta = 50$ のときでは, 高精度な検索ができているが, 比較回数が多くなってしまっている．改良手法でも, 比較回数を少なくするためには, パラメータ δ の値をうまく設定する必要がある．

5.　クラスタの階層化による索引構造

類似度計算回数を削減する手法について考察する．前節の検索では, 代表点への検索に「クラスタ数」回要してるため, 類似度計算回数の削減には, クラスタ数を削減しかつ高い再現率を維持するようなデータ構造が必要となる．本節ではクラスタの階層化を行い, 作成された上位層と下位層のクラスタの情報を用いることで, 計算回数を削減することを試みる．マージ処理の閾値を δ_1 として改良手法によりクラスタを作成したとする．より大きな閾値 δ_2 を用いて, 作成したクラスタを更にマージする．

上位層に δ の大きなもの, 下位層に δ の小さなもので作成されたクラスタの情報を用いて, 階層的に検索を行う．階層化の手順を **Algorithm 4** に示す．

Algorithm 4 改良手法によるクラスタの階層化手順

手順 1(k-MNN グラフ作成) $\mathcal{D}$ 中の全てのデータ間で類似度を計算し, k-MNN グラフを作成する．

手順 2(極大クリーク抽出) k-MNN グラフ中の全ての極大クリークを抽出する．得られた極大クリークを初期クラスタ集合とする．

手順 3(下位層クラスタの作成) 初期クラスタ集合に対し, マージの閾値を $\delta = \delta_1$ とし, 式 (3) を用いてマージ処理を行う．結果として得られるクラスタを下位層クラスタと呼ぶ．

手順 4(上位層クラスタの作成) 手順 3 で作成した下位層クラスタに対し, マージの閾値を $\delta = \delta_2$ とし, 式 (3) を用いてマージ処理を行う．結果として得られるクラスタを上位層クラスタと呼ぶ．

表 3　階層化により作成されるクラスタ ($k = 5, \delta_1 = 100$)

δ_2	クラスタ数	サイズ平均	サイズ分散	最大サイズ
7500	3491	12.10	147.79	183
10000	2780	15.17	245.39	310
15000	1900	22.15	487.36	274
20000	1272	33.05	1034.87	366
25000	834	50.24	2429.66	554
30000	455	91.77	7854.22	913

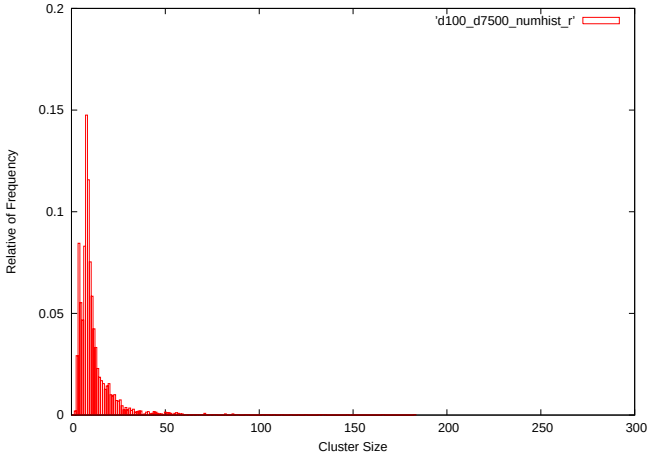


図 10　クラスタサイズの相対頻度分布 ($k = 5, \delta_1 = 100, \delta_2 = 7500$)

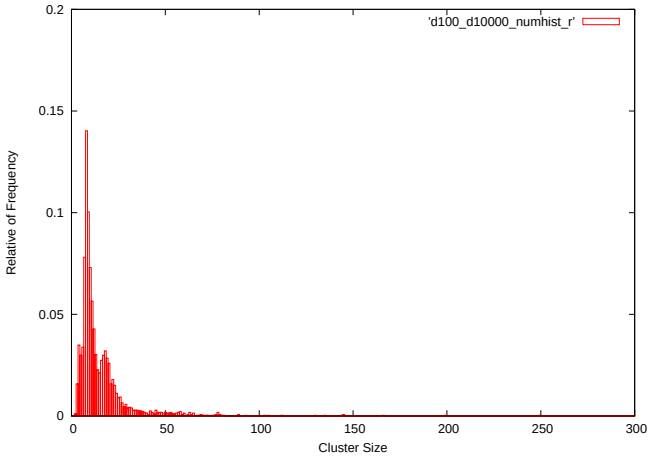


図 11　クラスタサイズの相対頻度分布 ($k = 5, \delta_1 = 100, \delta_2 = 10000$)

$k = 5, \delta_1 = 100$ で作成されたクラスタを入力とし, 上位層クラスタ作成時の閾値を $\delta_2 = 7500,\ 10000,\ 15000, 20000, 25000, 30000$ として階層化をそれぞれ行った. 作成される上位層クラスタの個数 (クラスタ数), 1 つのクラスタに含まれる領域画像数の平均, 分散, 最大 (サイズ平均, サイズ分散, 最大サイズ), を表 3 に示す.

また, $\delta_2 = 7500, 10000$ とし作成される上位層クラスタのクラスタサイズの相対頻度分布を図 10, 11 に示す.

5.1　上位層での検索実験

作成された上位層クラスタを用い, 検索実験を行う. 下位層クラスタの閾値を $\delta_1 = 100$ に固定して, 上位層クラスタの閾値を $\delta_2 = 7500, 10000, 15000, 20000$ とし, 作成したそれぞれのクラスタに対し, 4 節で用いた 1000 枚のクエリに対し **Algorithm 3** を適用し, 検索実験を行った. 出力する画像数を $l = 5$, $l = 20$, 選択クラスタ数 m の範囲を 1 から 300 とし, 全クエリに対する類似度計算回数の平均, 及び再現率の平均を求めたものを図 12, 13 に示す.

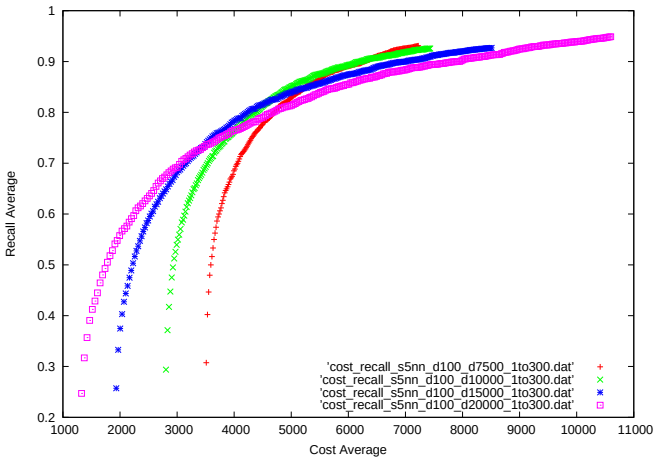


図 12　上位層クラスタを用いた検索 ($k = 5, \delta_1 = 100, l = 5$)

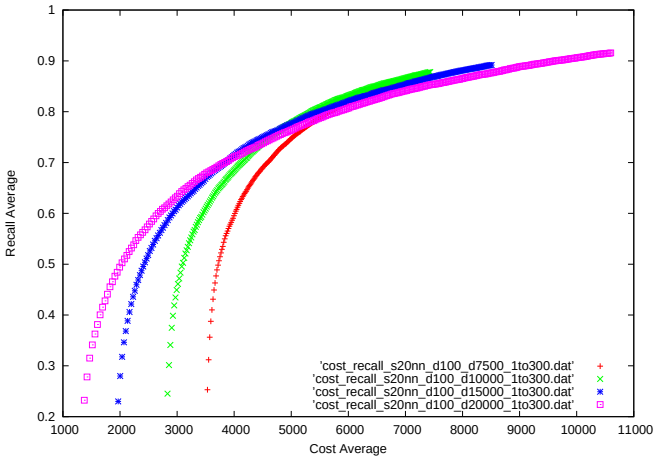


図 13　上位層クラスタを用いた検索 ($k = 5, \delta_1 = 100, l = 20$)

図 12, 13 の横軸は類似度計算回数の平均, 縦軸は再現率の平均を表す. 紫, 青, 緑, 赤の点はそれぞれ, 階層化手法により, $\delta_2 = 7500, 10000, 15000, 20000$ でマージしたクラスタを示す. 全体的に精度が若干低くなるが, クラスタ数の削減により類似度計算回数が抑えられている. 上位層のマージの閾値である δ_2 が大きくなるほどカーブがなだらかになる. すなわち, 再現率向上のために多くの類似度計算を要する. コストの観点から考えると, どのクラスタを用いるのが良いかは要求する再現率に依存する. 出力画像数が $l = 5$ で要求される再現率が 9 割の場合, $\delta_2 = 10000$ としたときの階層化クラスタを用い, $m = 209$ とすることで, 最も少ない類似度計算回数 (6213.78 回) で 9 割を達成できる. データベース全体 (34000) に占める割合はおよそ 18.27%程度であり, 計算回数が大幅に削減されている. 次節では下位層クラスタを用いることにより, さらに計算回数を削減する手法について述べる.

5.2　下位層クラスタを用いた検索実験

クラスタ数を減らすと, クラスタサイズが全体的に大きくな

Algorithm 5 階層化を利用した検索アルゴリズム

手順 1(上位層クラスタの選択) 作成した各上位層クラスタ $\mathcal{C}_i(i = 1, \ldots, M)$ に対し, クエリと代表点 p_i との類似度を計算し, 類似度の大きいものから上位 m 以内のクラスタを選択する. m を選択クラスタ数と呼ぶ.

手順 2(下位層の代表点による絞り込み) 手順 1 で選択した各クラスタ $\mathcal{C}_i$ について, 次の処理を実行する. $\mathcal{C}_i$ に含まれる下位層クラスタの各代表点とクエリとの $simRank$ を計算し, 値が α 以下ならば, $\mathcal{C}_i$ の要素を検索リスト A に追加する.

手順 3(類似画像の出力) A に含まれる全ての要素とクエリとの類似度を計算し, 上位 l 個を出力結果とする.

表 4 上位層と下位層を用いた検索結果 ($k = 5, \delta_1 = 100, \delta_2 = 7500, m = 300, l = 5$)

α	類似度計算回数の平均	再現率の平均
500	5420.97	0.8920
1000	5682.27	0.9108
2000	6091.60	0.9234
3000	6376.88	0.9268
4000	6594.87	0.9284
5000	6770.02	0.9292
6000	6907.42	0.9298
7000	7012.24	0.9300
8000	7095.66	0.9306
9000	7164.89	0.9308
10000	7222.28	0.9308

るため, クラスタを選択した際の選択リスト $|A|$ の値が大きくなる. これを削減するために下位層クラスタの情報を用い, 計算回数を削減することを考える.

選択した上位層クラスタ $\mathcal{C}_i$ とする. $\mathcal{C}_i$ に含まれる各下位層クラスタ $\mathcal{C}_j$ について, $\mathcal{C}_j$ の代表点とクエリとの $simRank$ が α 以下であるならば $\mathcal{C}_i$ を選択リスト $|A|$ に追加する.

検索の手順を **Algorithm 5** に示す.

各パラメータを $k = 5, \delta_1 = 100$, $\delta_2 = 7500$ として作成したクラスタ, 及び, $k = 5, \delta_1 = 100$, $\delta_2 = 10000$ として作成したクラスタを用いて実験を行った. 出力する画像数を $l = 5$, 上位層の選択クラスタ数を $m = 300$ として下位層を用いた検索実験を行った. 下位層のパラメータ α の値を変化させ, 類似度計算回数と再現率の平均を求めたものを表 4, 5 にそれぞれ示す. 下位層のクラスタを用いることにより, 再現率を維持しつつ, 計算回数が削減できているのがわかる. $\delta_2 = 7500$, $m = 300$, $\alpha = 1000$ では, 5682.27 回 (全体の 16.7%) の検索回数で, 再現率 91%を達成している. また, $\delta_2 = 10000$, $m = 300$, $\alpha = 1000$ では 5375.91 回 (全体の 15.8%) の検索回数で, 再現率 90%を達成している. α が 2000 以上のとき, それ以上値を大きくしても再現率に大きな差は見られない. これは, 下位層クラスタの代表点との $simRank$ が 2000 以上のものを枝切りしても, 再現率に大きな影響を与えないことを表している.

図 14, 15 に, 上位層と下位層を用いた検索 (赤点) と, 上位層のみを用いた検索 (緑点) の比較を示す. 赤点は, 上位層と下位層を用いた際の類似度計算回数と再現率の平均 (左から順に $\alpha =$

表 5 上位層と下位層を用いた検索結果 ($k = 5, \delta_1 = 100, \delta_2 = 10000, m = 300, l = 5$)

α	類似度計算回数の平均	再現率の平均
500	5112.03	0.8842
1000	5375.91	0.9034
2000	5818.82	0.9154
3000	6151.30	0.9196
4000	6413.98	0.9212
5000	6632.10	0.9226
6000	6816.02	0.9238
7000	6965.40	0.9246
8000	7085.34	0.9252
9000	7182.85	0.9256
10000	7266.40	0.9256

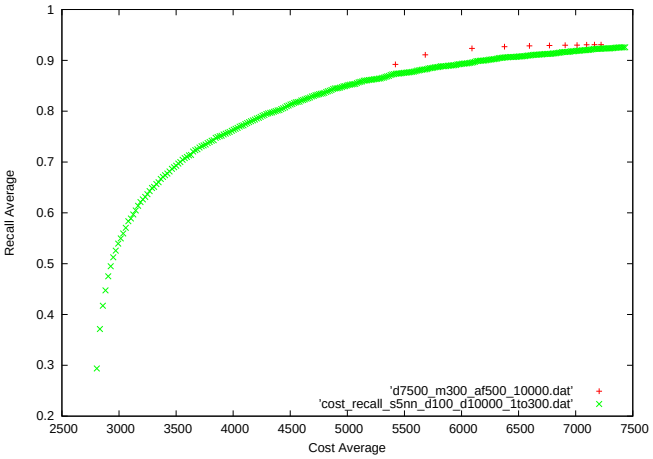


図 14 階層化検索の比較 ($k = 5, \delta_1 = 100, \delta_2 = 7500, m = 300, l = 5$)

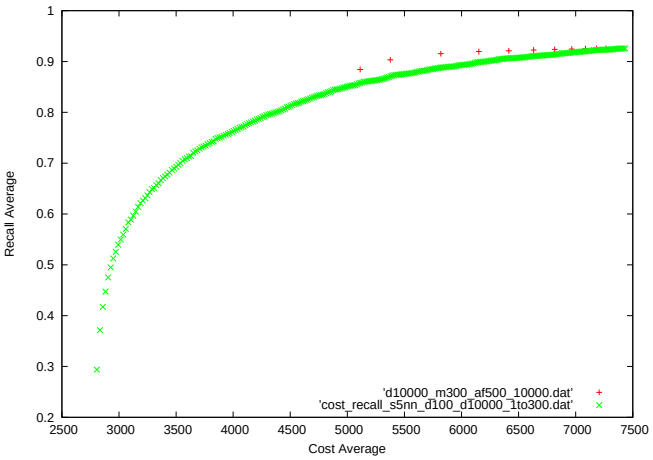


図 15 階層化検索の比較 ($k = 5, \delta_1 = 100, \delta_2 = 10000, m = 300, l = 5$)

$500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000$ を表す), 緑の点は選択クラスタ数 m の範囲を 1 から 300 とした際の上位層のみを用いた検索結果である. 図 14, 15 より, 少ない類似度比較回数で, 高い再現率を実現する検索が出来ているため, 本階層化手法を用いた検索は有効だと言える.

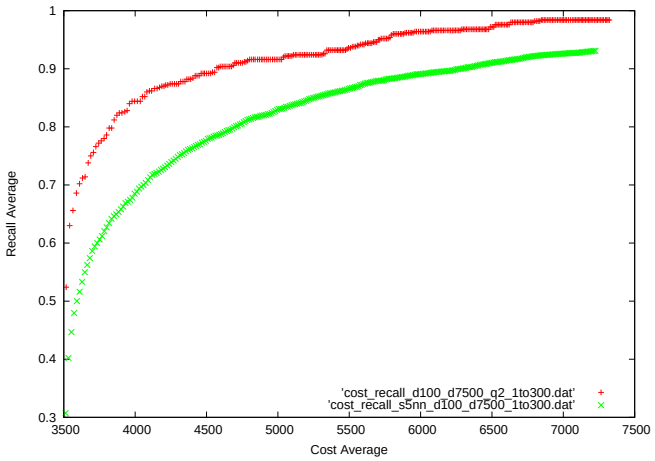


図 16　孤立点を除いたクエリ集合との比較 ($k = 5, \delta_1 = 100, \delta_2 = 7500, l = 5$)

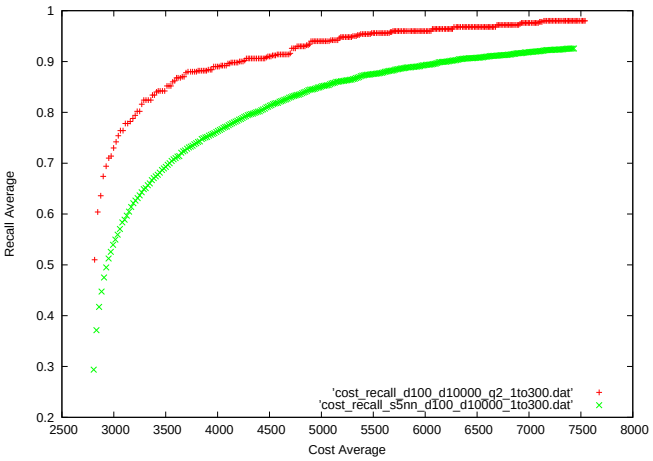


図 17　孤立点を除いたクエリ集合との比較 ($k = 5, \delta_1 = 100, \delta_2 = 10000, l = 5$)

6. クエリ集合を変更した場合の検索実験

本節では, データベース $\mathcal{D}$ の中にクエリと類似度の高いデータが複数存在する場合の類似検索について調べる. そこで, 類似度が高いデータが複数存在する画像のみからなるクエリ集合を新たに作成し検索実験を行う.

$k = 5$ で k-MNN グラフを作成し, 極大クリークを抽出した際に, 要素数が 3 以上の極大クリークに含まれる画像からランダムに 100 枚を選び, 新たなクエリ集合 (孤立点を除いたクエリ集合) とする.

$k = 5, \delta_1 = 100, \delta_2 = 7500, 10000$ とし作成されるそれぞれの上位層クラスタに対し, 孤立点を除いたクエリ集合および, 4 節で用いた 1000 枚のクエリ集合それぞれに対し, 検索実験を行い結果を比較する.

出力する画像数を $l = 5$, 選択クラスタ数 m の範囲を 1 から 300 とし, 作成したクエリの全クエリに対する類似度計算回数の平均, 及び再現率の平均を求めたものを図 16, 17 に示す. 赤点は本節で作成した孤立点を除いたクエリ集合を用いた結果, 緑点は 4 節で用いた 1000 枚のクエリ集合を用いた結果を表す.

いずれのクラスタにおいても, 本節で作成した新たなクエリ集合のほうが良い結果となっている. 選択クラスタ数を $m = 300$ とすると再現率は 0.98 以上となり, 全検索とほとんど変わらない結果となっているのがわかる.

以上の結果は, データベース内に似ている画像が含まれているクエリの場合, 本検索手法を用いることにより, 非常に高い再現率での検索が行えることを示している.

7. ま と め

本稿では金城らが提案したクラスタリング手法と検索手法の改良を行った. 作成されるクラスタをランクの観点から分析し, クラスタにノイズとなる要素が含むことを確認した. この問題に対し, 我々はクラスタ内に含まれる画像の $simRank$ 最大値に着目し, マージ処理の改良を行った. 改良手法の評価のため, 34000 枚の領域画像データを用いてクラスタ作成実験を行い, 従来手法で作成されるクラスタとの孤立点の個数と $simRank$ 最大値について, 従来手法との比較を行った. 更に, 改良手法によって作成したクラスタを用いて検索実験を行った. 実験の結果から改良手法を用いることにより, 高精度な検索を実現できることを示した.

クラスタを用いた検索を更に高速に行う手法として, 階層化による索引構造を提案した. 実験の結果から, 類似度計算回数は大幅に削減され, 高速かつ高精度な検索が実現できることを示した.

また, 索引構造の特徴を調べるため, クエリ集合を変更した場合の検索実験を行った. 実験の結果から, データベース内に似ている画像が含まれているクエリの場合, 非常に高い精度での検索が行えることを示した.

文　献

[1] A.K.Jain, M.N.Murty, and P.J.Flynn. Dataclustering: A review. ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.

[2] Zaher Dawy, Bernhard Goebel, Joachim Hagenauer, Christophe Andreoliand, Thomas Meitinger, and Jakob C.Mueller. Gene mapping and marker clustering using shannon's mutual information. IEEE/ACM Trans Comput. Biol. Bioinformatics (TCBB), Vol. 3, No. 1, pp.47-56, 2006

[3] K.Chidananda Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern Reconition, Vol. 10, No. 2, pp. 105-112, 1978.

[4] R.A.Jarvis and Edward A.Ratrick. Clustering using a similarity measure based on shared near neighbors. IEEE Transactions on Computers, Vol. C-22, No. 11, pp. 1025-1034, 1973.

[5] S.Khy, Y.Ishikawa, and H.Kitagawa. Novelty-based incremental document clustering for on-line documents. In Proc. of International Workshop on Challenges in Web Information Retrieval and Integration(WIRI 2006), 2006.

[6] 金城 賀真, "領域分割に基づく類似画像検索の高速化", 電気通信大学大学院情報システム学研究科修士論文, 2008.

[7] 金城 賀真, 小早川 倫広, 星 守, 大森 匡, 距離の公理を満たさない類似度を用いたクラスタリング手法, DEWS2008, E3-1, 2008.

[8] 山本 敦, 小早川 倫広, 星 守, 大森 匡, 画像の領域分割に基づく類似画像検索, 情報処理学会論文誌, Vol. TOD 35, No. 46, pp. 93-105, 2007.