

情報紋 InfoPrint を用いたマルチメディア情報の類似検索手法

並 河 祐 貴^{†1}

川 越 恭 二^{†2}

様々なマルチメディア情報から構成されるコンテンツが広域にわたり日々生成・蓄積されている。蓄積されたマルチメディア情報から必要な情報を取り出すためには、一般に多大な検索時間・計算量を必要とする。これは、マルチメディア情報を直接に検索することができず多次元特徴ベクトルによる比較を行うためである。これらの問題を解決するため、本稿では、マルチメディア情報の持つ多次元特徴ベクトルから生成した画像データ（情報紋，InfoPrint と呼ぶ）の特徴を用いることで大量のマルチメディア情報の集合に対する類似検索処理の効率を向上させる手法の提案を行う。

Similarity Search Method for Multimedia Content using Images, called InfoPrint, derived from its Multi-dimensional data

YUUKI NAMIKAWA^{†1} and KYOJI KAWAGOE^{†2}

Various multimedia data have been generated and stored on internet rapidly. It takes considerable time to search a desirable data among huge amount of such multimedia data and to process such similarity searching because the multimedia data is represented as a point in a multi-dimensional space. In order to solve the inefficiency of the similarity search, we propose a new method in which InfoPrint, a graphical image derived from multi-dimensional data of multimedia content, is used for decreasing the number of similar candidates.

1. はじめに

近年，ネットワークインフラの急速な発展と，インターネット技術のめざましい発達により，人々の持つ情報や知識の共有に大きく貢献している。インターネットを通じて得られる大量の情報は，巨大なデータベースとして日々蓄積されている。テキスト・画像・音楽・映像などの様々な情報は，マルチメディアコンテンツとしてインターネットに多数存在している。このため，日々膨大に生成・蓄積されていくマルチメディア Web ページ集合に対して，必要な情報を効率よく取り出す必要性が高まってきている。

また，マルチメディアコンテンツを構成している様々な種類のマルチメディア情報を検索するには，多次元の特徴ベクトル同士で比較することが多い⁸⁾。しかし，インターネット上の大量のデ

ータ集合を対象とした際には，検索に多大な計算量が必要となり，検索時間が増大するだけでなく検索処理がシステムに大きな影響を与えかねない。このため，処理負荷を軽減させて効率の良い検索手法を用いる必要性がある。

そこで，本稿では，ネットワーク上に広域散在している大量のマルチメディアコンテンツから，クエリとして与えた Web ページ集合と類似したコンテンツを効率的に抜き出すための新しい類似検索手法を提案する。提案する方法は，複数のマルチメディア Web ページをクエリとして与え，その複数のページからメディアごとに特徴量を抽出し，その特徴量を効率的に用いて類似検索処理を行う方法である。特に，類似検索の前処理として，マルチメディア情報の持つ多次元特徴ベクトルから，情報紋として 1 枚の画像を生成し利用する点に特徴がある。この生成画像の特徴をインデックスとしてフィルタリングに利用することで，明らかに類似していない大量の Web ページ

^{†1} 立命館大学大学院 理工学研究科

Graduate School of Science and Engineering, Ritsumeikan Univ.

^{†2} 立命館大学 理工学部

School of Science and Engineering, Ritsumeikan Univ.

を排除し、計算量の削減を図ることが可能となる。

提案するマルチメディア情報類似検索手法が適用可能な代表的な応用は、インターネット上で不正に公開されている著作物(デジタルコンテンツ、マルチメディアデータ等)の探索である。この応用では、本手法を適用することで、正規のデジタルコンテンツに類似した不正コンテンツを効率的に類似検索できるものとする。また、Web サイト内の一部の情報から、Web サイト自体の検索を行う高度な検索エンジンとしての応用も、提案する方法によって可能であるとする。

本稿の構成は以下のとおりである。2 章では、マルチメディア情報検索の概要について述べ、3 章では、本稿で提案する、情報紋 InfoPrint を用いた類似検索手法について述べる。4 章では、シミュレーションによる実験とその結果について述べ、最後に 5 章では、まとめ及び今後の課題について述べる。

2. マルチメディア情報の検索

マルチメディア情報の検索は、メタデータに基づいた検索と、内容に基づいた検索の 2 つに大きく分類できる⁸⁾。

メタデータに基づいた検索では、マルチメディア情報に対して意味付けされた情報や内容の要約情報など、そのデータがどういったものであるかを表現したメタデータを基にした検索である。一方、内容に基づいた検索は、マルチメディア情報そのものの内容の比較に基づいた検索である。本稿では、Web ページ集合をクエリとし、マルチメディア情報の持つ特徴を基に検索を行うため、以降、内容検索を対象としている。

このようなマルチメディア情報の検索を行うには以下の 2 種類の考え方が存在する⁹⁾。最初の考え方は、大量のデータ集合の中から、いかに効率よく利用者の望む情報を探し出せるかという点である。2 番目の考え方は、マルチメディア情報自体のサイズも大きいので、いかに少ない計算量で、データ同士の比較が行えるか、という点である。これらの 2 つの考え方を実現するために、これまでに様々な方式が提案されている¹⁾⁴⁾⁶⁾。

マルチメディア情報の検索を行う上で、各データ間の類似度は通常、特徴量空間における距離を用いて判定する。類似度とは、比較するデータがどれだけ似た特徴を持っているかを示した値で

あり、距離が近いほど類似度が高く、距離が遠いほど類似度は低いと判定する。距離の計算方法には、本稿では一般的に広く使われているユークリッド距離 (L_2 距離) を用いる。

ここでユークリッド距離とは、検索を行う上で比較する 2 つのデータが持つベクトルの各成分の差の 2 乗を求め、その和を用いて 2 つのベクトルの距離を評価する尺度である。2 つの任意の n 次元ベクトル、 $A=\{a_1, a_2, \dots, a_n\}$ と $B=\{b_1, b_2, \dots, b_n\}$ に対して、 A と B との間のユークリッド距離を E とすると、

$$E = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

と表すことができる。成分の差を 2 乗することによって、成分の差を強調することができる。このユークリッド距離関数を用いたマルチメディア情報の類似検索処理を行うことで、精度の高い検索を行うことが可能となる。

ユークリッド距離関数を用いた類似検索処理では、検索を行うデータと検索対象となるデータのそれぞれの各要素を、検索処理時に直接比較を行うため、あらかじめインデックスを作成しておくことが難しい。また、大量のデータ集合を対象とした場合、比較時に発生する莫大な計算量が問題となる。このため、類似検索処理の際の計算量を削減するために、あらかじめ大量のデータに対して絞り込みをしておく必要がある。

3. 情報紋 InfoPrint を用いた類似検索

本稿では、2 章で述べた検索処理時における計算量の増大化を抑えるために、情報紋 InfoPrint を用いた効率的な検索手法を提案する。本章では提案方式について詳しく述べる。

3.1 提案方式の基本的考え方

一般に、テキストや画像、音楽、映像等のメディア情報の特徴量空間は、多次元であると考えられる。このため、インデックスとして格納する Web ページ数が増加するほど、比較の際の計算量が膨大となり、検索処理やフィルタリングに時間がかかるという問題が発生する。そこで、本方式では、インデックスとして扱う Web ページ集合の多次元特徴ベクトルを、1 枚の情報紋となる 2 次元画像データ (以下、InfoPrint と記す) に変換し、

その画像データ InfoPrint の特徴をインデックスとして用いることを提案する。

InfoPrint に変換する理由を以下に列挙する。

- 1) 量的に均質な情報に圧縮できる
- 2) 多次元の情報を圧縮させることが可能
- 3) 画像の類似検索手法の適用が可能
- 4) 人間にもビジュアルである

こうして作成した InfoPrint の特徴をインデックスとして用いて、Web ページ集合の検索処理におけるフィルタリングとして利用する。これにより、類似検索を行う前処理として、検索条件と完全にかけ離れて明らかに類似していない Web ページ集合を排除することができ、検索処理の効率化を図ることが可能であると考える。

3.2 提案方式の処理フロー

3.1 の基本的考え方に基づいた提案方式の処理の流れを説明する。以下に示す手順で、マルチメディア Web ページの類似検索処理を行う。図 1 に提案方式の処理の流れを示す。

Step1 最初に、利用者から、検索対象とするコンテンツとして複数の Web ページを取得する。

Step2 検索対象として与えられた複数の Web ページから、それぞれの Web ページに含まれているメディアごとに解析を行い、特徴量を抽出する。

Step3 Step2 で抽出された特徴量を、メディアごとにマッピングする。例えば、テキストならテキストを対象とした n 次元の特徴量空間に対して、与えられたページごとにマッピングを行う。

Step4 Step1 で与えられた各ページの点からを含むような最小の矩形領域を作成する。例えば、クエリとして 3 点の Web ページが得られたとすると、図 2 のようにページごとに 3 つの点を n 次元の特徴量空間上に配置し、最も外側の点に外接するように矩形を求める。この矩形により、得られた Web ページ集合を管理する。同時に、その Web ページ集合の点のそれぞれの重心 G も算出しておく。

Step5 Step4 で作成された多次元インデックスを、必要であれば次元縮小を行った上で、InfoPrint(画像データ)に変換し、

その画像の特徴を用いてフィルタリングを行う。これにより明らかに類似していない Web ページ集合をあらかじめ類似検索の前処理によって排除する。尚、InfoPrint については 3.3 で、次元縮小については 3.4 で詳しく述べる。

Step6 InfoPrint によるフィルター処理の後、Step4 で作成した領域情報や重心 G をインデックスとし、特徴量空間上でユークリッド距離関数を用いて、領域間の距離を測ることにより、類似検索処理を行う。

Step7 得られた Web ページ集合間の類似性を算出し、類似性の高い順序に利用者に提示する。

3.3 InfoPrint

本研究では、マルチメディア情報の多次元特徴ベクトルに対して画像変換を行い、その画像の特

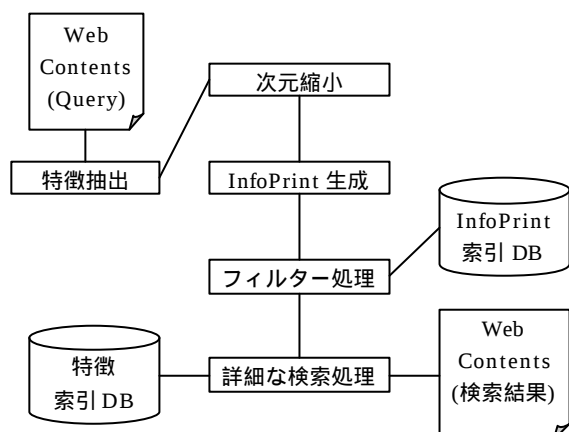


図 1 提案方式の処理の流れ

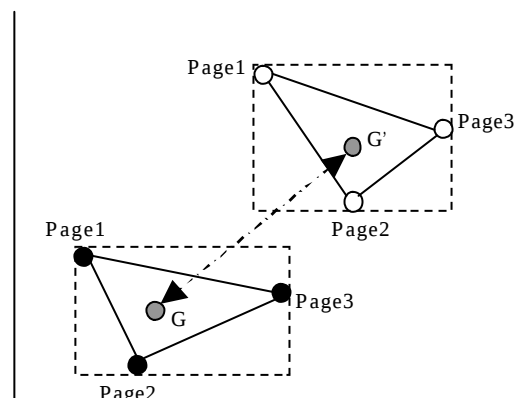


図 2 特徴量空間への配置例

徴を用いてフィルタリングを行う。描画する図形としては、マンデルブローフラクタル²⁾を用いた。理由として、後の類似画像の検索であるフィルタ処理を考慮した際、マンデルブローフラクタルが特徴的な図形であり、画像の比較に適していると考えられるからである。

3.3.1 マンデルブロー集合

マンデルブロー集合は一般的に次の反復式で定義されている⁷⁾。

$$Z_{n+1} = Z_n^2 + C \quad (1)$$

上記の式において、

$$Z_n = x_n + y_n i$$

$$C = a + bi$$

とする。式(1)の反復式は定数 C の値によって、 Z_n が 0 に近づいたり、無限に大きくなったりする。また、反復条件を設け、複素平面上の点 Z_n の原点(初期値)からの距離を用いて収束か発散かを判定し、発散なら、反復回数に応じて、その座標領域の点に対して色付けを行う。この定数 C の点を連続的に変化させて、マンデルブロー集合の模様を描画することが可能である。図 3 は実際にマンデルブロー集合を描画した例である。

3.3.2 多次元データからの適用方法

マンデルブローフラクタルに変換するために、複素平面上の点の指定や、図形の一部と拡大率を指定する。フラクタルは図形の一部を任意に拡大して表示することが可能であり、複数回拡大しても空にはならず、無限の微細構造をもつといった特徴がある。この特徴を利用し、拡大する都度、拡大に必要なパラメータを与えていく。これにより、描画の際に無限の個数の変数を扱うことが可能である。尚、拡大のために必要なパラメータとしては、拡大部分の座標(x, y)や拡大率がある。

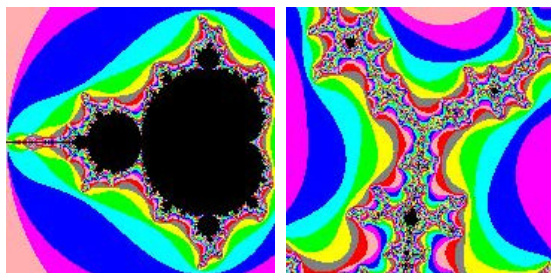


図 3 マンデルブローフラクタル

フラクタル図形を描画する際、数千回の反復計算を行い、さらに拡大することで、数千次元の特徴ベクトルが扱うことが可能である。しかし、実際に数千回反復計算を行って描画された画像をインデックスとしてフィルタリングを行った場合、本当に不必要な情報だけを排除できるかどうかは疑問である。

そこで、フラクタル図形への変換の際に扱える変数の個数以上の次元数を持つ多次元特徴ベクトルを扱うために、InfoPrint への変換の前処理として次元縮小を考える。次元縮小を可能とする方式として、KL 展開(Karhunen-Loeve 展開)³⁾を用いることとする。KL 展開は、そのままでは扱うことの困難な高次元のデータに対して情報損失を抑えつつ次元の圧縮を行う方式として通常、用いられる方法である。KL 展開については 3.4 で詳しく述べる。

3.4 KL 展開 (Karhunen-Loeve 展開)

KL 展開(Karhunen-Loeve 展開)³⁾は、多次元特徴ベクトルの分布全体が持つ情報に対して、なるべく情報量を落とさず、最大限に情報を反映できるように特徴量空間の次元を削減する手法として知られており、数学的には主成分分析と等価である。

KL 展開は大きく分けて 3 つの処理から構成されている⁵⁾。まず、入力された多次元データの共分散行列 Σ を計算し、次に Σ の固有ベクトル E を求め、最後に E から基底を求め、次元縮小を行う。以下、それぞれの処理の概要を述べる。

- 1) まず、与えられた多次元データの特徴空間における共分散行列を計算する。共分散は以下の式で求めることが出来る。

$$a_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

ここで、 $\bar{x}_i$ と $\bar{x}_j$ はデータの第 i, j 番目の要

素の平均を、 n はデータの総数を、 a_{ij} は第 i 番目の要素の間の共分散を表す。

- 2) 次に、得られた共分散行列の固有値を求め、その固有値の大きいものから m 個の $\lambda_1, \lambda_2, \dots, \lambda_m$ を選び、各固有値に対応する固有ベクトル $y_1, y_2, \dots, y_m$ を求める。ここで、行列 Σ の固有値、固有ベクトルとは、

$\lambda_i y_i = \sum y_i \quad (i = 1, 2, \dots, m)$ という関係を

満たすものである。

- 3) 最後に得られた固有ベクトルから基底を求め、次元縮小を行う。基の特徴量空間を m 次元の部分空間に射影するためには固有ベクトルから固有値の大きい順に m 個の固有ベクトルを取り出し、それを基底としてデータを部分空間に射影する。

本稿では、与えられた多次元のデータに対して、InfoPrint の生成に必要な次元数までの圧縮手法として、KL 展開の手法を用いる。

3.5 類似検索処理

提案する方法は、InfoPrint を用いて類似していないデータを以降の処理対象からはずすことで効率的な類似検索を可能とする方法である。類似検索対象となったデータは、その多次元特徴ベクトルで構成される特徴量空間上にマッピングされる。具体的には、マッピングされた重心及び領域情報に対して、ユークリッド距離関数を用いて各領域情報の比較を行うことにより、マルチメディア情報の類似検索を行う。

すなわち、クエリのデータに対応した重心と領域を用いて、その重心と近くの重心を持ち、しかも領域のオーバーラップが多いようなデータを順にあらかじめ定められた条件を満たす類似データを求める。実際にこれらのデータが類似しているかどうかについては、個々の Web ページ集合間の類似性を算出することで行う。

4. シミュレーション実験

これまでに説明した本方式の有効性を確認するためにシミュレーションプログラムを作成し、InfoPrint の有効性と、3 種類の分布に従ったデータに対してフィルター処理を行い、検索精度を測定する実験を行った。

4.1 実験 1

InfoPrint の生成は、多次元の特徴ベクトル、すなわち複数の変数を基にして行われる。そこで、実験 1 では、その複数の各変数について、変数の値を少しずつ変化させた InfoPrint を比較することによって、InfoPrint 自体の類似性と変数の差の関係を調べた。

まず、ランダムな複数の変数を与えてフラクタル図形を描画させた画像(InfoPrint)をオリジナルとし、そのオリジナルの InfoPrint を生成する際に用いられた複数の変数の内の 1 つの変数の値を変化させて描画した InfoPrint 20 枚と、基のオリジナルの InfoPrint との画像の一致率および変数の差の関係をグラフにした。

図 4 ではオリジナルイメージを G_1 、 G_2 、および G_3 として、それぞれのイメージを構成している変数 P_1 の値から、差 α だけ変化させて生成したイメージ G_1' 、 G_2' 、および G_3' と、オリジナルイメージとの画像の一致率とその 3 つのグラフの平均を示したグラフとなっている。

図 4 より、変数 P_1 の値は、差が大きくなるにつれて、画像の一致率が低下していくことを示している。これはオリジナルの InfoPrint と比較する InfoPrint の特徴ベクトルの特定の要素の値の差が大きくなれば、InfoPrint の類似性が低くなることを示している。

また、3.3.2 で記した、マンデルブローフラクタルである InfoPrint を生成する際に必要となる最低限の変数においても、同様の実験を行い、その全ての変数において、オリジナルとなる InfoPrint の変数との差が大きくなるにつれて画像の一致率が下がっていく、つまり類似性が低くなるといった傾向を確認した。

この実験結果より、Faloutsos による no false dismissal の特性⁹⁾を持つことがうかがえる。

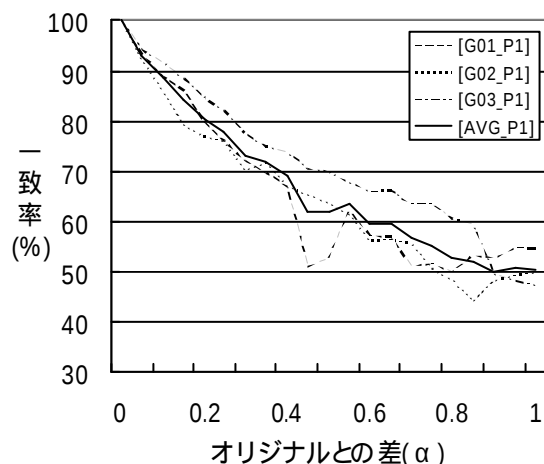


図 4 実験 1 の結果

4.2 実験2

実験2では、本稿で提案する InfoPrint を用いた類似検索手法の検索精度を調べるために、再現率、適合率、及び再現率・適合率の結果の幾何平均の測定を行った。

まず、検索を行うデータ集合に対して、各領域情報の特徴ベクトルを用いて、あらかじめユークリッド距離関数を用いた検索を行っておき、ユークリッド距離の値の小さいデータの上位10件を求めておいた。つまりデータ集合の中からクエリと比較して類似性の高いデータ10件をあらかじめ求めておく。その10件のデータを正解集合としておき、本検索手法で検索を行った際、検索結果にどれだけ正解集合のデータが含まれているかを調べた。検索精度の指標として再現率と適合率、及びそれらの幾何平均を用いる。再現率及び適合率は次式より求める。

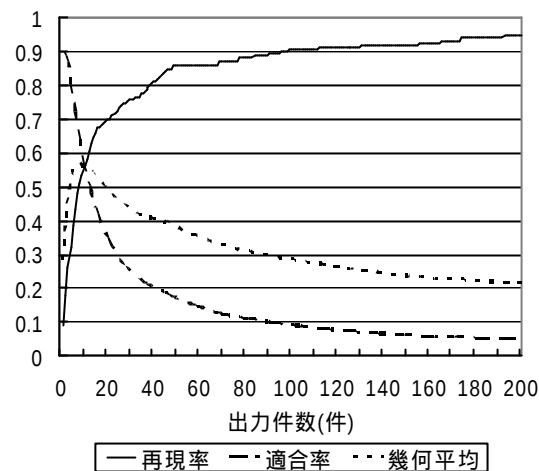
$$\text{再現率} = \frac{\text{検索結果に含まれる正解の数}}{\text{全正解数}}$$

$$\text{適合率} = \frac{\text{検索結果に含まれる正解の数}}{\text{検索結果の数}}$$

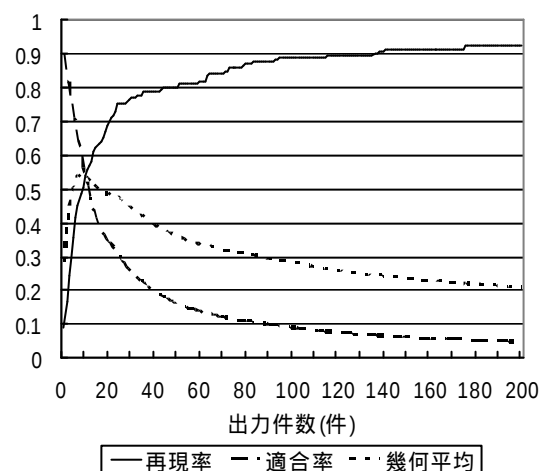
また、幾何平均は、再現率と適合率の積の2乗根をとったものである。

検索対象とするデータは、3種類の200次元の特徴ベクトルを持つ擬似データを用意した。ランダムな分布に従った数値集合のデータを1200件、正規分布(ガウス分布)に従った数値集合のデータを300件、多次元実データの分布¹⁾に従った数値集合データを300件、の3種類である。それぞれの実験では、検索回数を20回とし、正解集合の個数を10件とした。また、あらかじめ200次元の特徴ベクトルからKL展開を用いて40次元まで次元数の削減を行い、その後、InfoPrintに変換した。InfoPrintに変換する際には、生成する時に必要となる各変数をInfoPrintがフィルタリングとして有効である範囲への正規化を行った上で変換を行った。その後、InfoPrintの特徴を用いて類似検索処理を行った。

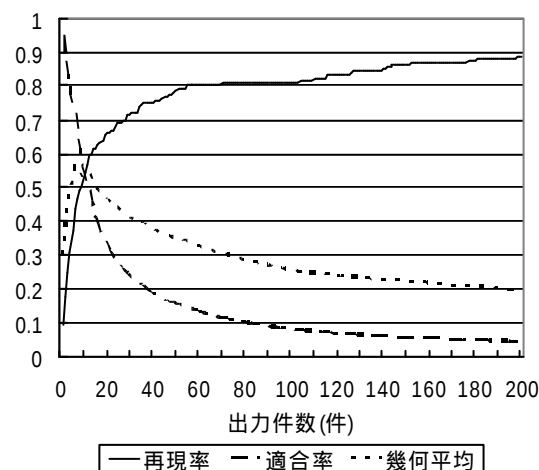
最初に、ランダムな分布に従ったデータに対して、本稿で提案した InfoPrint を用いた類似検索処理を行った。検索対象とするデータ数を300件、600件、1200件と変化させて、それぞれについて検索を行い、再現率、適合率、及び幾何平均の測定を行った。この時の測定した各指標の推移を図5に示す。



(a) 検索対象となるデータ数が 300 件の場合



(b) 検索対象となるデータ数が 600 件の場合



(c) 検索対象となるデータ数が 1200 件の場合

図5 各データ数における再現率・適合率・幾何平均の推移

図 5 より、どのデータ件数においても、検索結果の出力件数がおよそ 10 件付近で幾何平均が最大となる。また、幾何平均が 0.4 以上の点に対しては、再現率及び適合率の双方が比較的良好な値を示しているため、幾何平均が高くなる 0.4 以上となる点のデータ件数までのフィルタリングに対して、優れた性能を示しているといえる。

また、例えば再現率が 85%となるデータ数までの絞込みを考えた場合、図 5 より検索対象となるデータ数が 300 件の場合には 45 件付近までの 15.0%、データ数が 600 件の場合は 70 件付近までの 13.3%、データ数が 1200 件の場合には 140 件付近までの 11.7%への絞込みが可能である。これより、検索対象となるデータ数が増えるほど、全データ件数に対して、絞り込める件数の割合が小さくなっていることがわかる。すなわち、検索対象となるデータが増えれば、InfoPrint を用いたフィルター処理により排除できるデータの割合が増加することを意味する。この結果、InfoPrint を用いたフィルター処理は、大量のデータ集合を対象とした類似検索の処理効率に大きく貢献できると考える。

次に、より現実的な分布を持つデータに対する有効性を示すために、正規分布に従った値を持つデータ 300 件と、多次元実データの分布に従った値を持つデータ 300 件の 2 つに対しても、ランダムな分布に従った値を持つデータと同様の実験を行った。3 種類のデータの比較実験では、それぞれ検索対象となるデータは 300 件、検索回数は 20 回とし、正解集合の個数を 10 件として、それぞれ検索を行った。この実験による各々の分布に従ったデータの持つ再現率、適合率、及び幾何平均の推移を図 6 に示す。また、各々の分布に従ったデータが最も大きい幾何平均をとる時の再現率・適合率を表 1 に示す。

図 5(a)、図 6、及び表 1 より、各々の分布に従ったデータの再現率、適合率、及び幾何平均は、3 種類とも類似した推移を示した。この実験の結果、ランダムな分布に従ったデータだけでなく、特徴的な分布に従ったデータや、実データに近いデータに対しても、本方式が有効であると考えられる。

5. おわりに

本稿では、マルチメディア情報に対する検索効率を向上させるため、情報紋 InfoPrint の特徴を

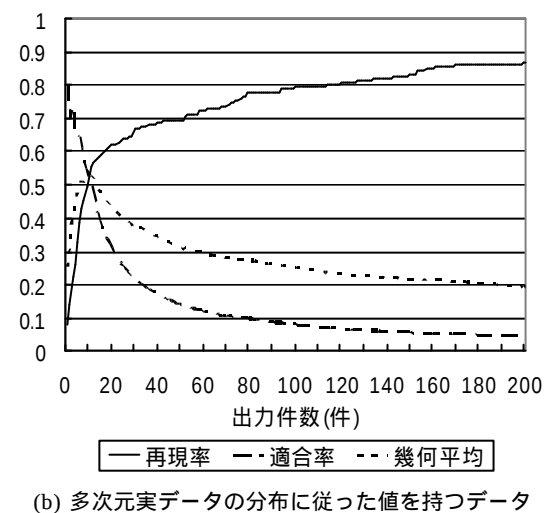
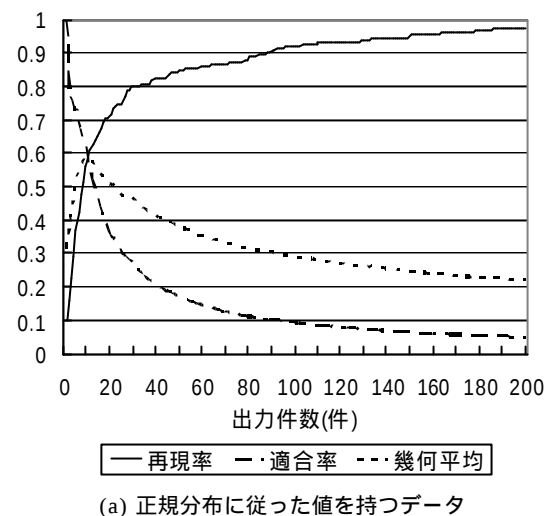


図 6 各分布に従った値を持つデータによる再現率・適合率・幾何平均の推移

表 1 各分布に従ったデータの再現率・適合率

	再現率	適合率
ランダムデータ	53.3%	59.2%
正規分布データ	56.0%	62.2%
多次元実データ	55.3%	50.3%

用いた類似検索手法を提案した。提案した方式では、大量のデータ集合に対して、ユークリッド距離を用いた検索処理の前に、InfoPrint の特徴を用いて類似していない情報を十分に排除することにより、効率のよい検索を行うことが可能である。また、シミュレーション実験により、InfoPrint の有効性を示した。

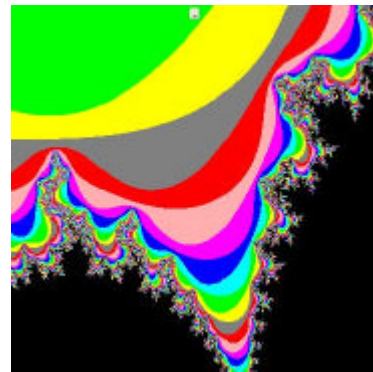
提案方式における今後の課題を以下に示す。

- より大量のデータに対する実験
 - マルチメディア情報から構成される実 Web データを用いた評価実験
 - フラクタル図形以外の InfoPrint の検討とその検索精度に関する実験
- 等を行う予定である。

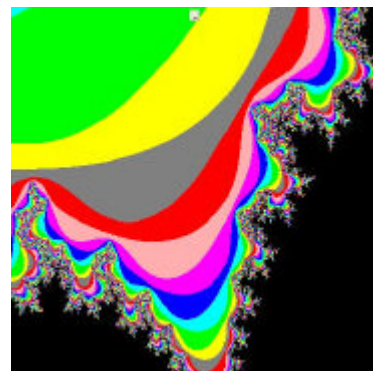
参 考 文 献

- 1) 安 際元, 古瀬 一隆, 陳 漢雄, 石川 雅弘, 大保 信夫 : 凸多面体を用いた次元縮小法と高次元索引機構, 情報処理学会論文誌, Vol.43, No.SIG2, pp.168-177 (2002)
- 2) 鈴木 優, 波多野 賢治, 吉川 正俊, 植村 俊亮 : 複数のメディアで構成された電子文書の検索手法, 情報処理学会研究報告, DBS-122-17, pp.129-135 (2000)
- 3) Fukunaga, K. : Statistical Pattern Recognition, Academic Press (1990)
- 4) Christos Faloutsos, King-Ip (David) Lin : FastMap: A Fast Algorithm for Indexing, Data Mining and Visualization of Traditional and Multimedia Datasets, Proc. 1995 ACM SIGMOD International Conference on Management of Data, pp.163-174 (1995)
- 5) 長尾 真, 松山 隆司, 杉本 晃宏, 佐藤 理史, 麻生 英樹 : 岩波講座 マルチメディア情報学 2 情報の組織化, 岩波書店 (2000)
- 6) 西尾 章治郎, 田中 克己, 上原 邦明, 有木 康雄, 加藤 俊一, 河野 浩之 : 岩波講座 マルチメディア情報学 8 情報の構造化と検索, 岩波書店 (2000)
- 7) Hans Lauwerier : Fractals: Endlessly Repeated Geometrical Figures, Maruzen Co. (1996)
- 8) 吉川 正俊, 植村 俊亮 : マルチメディアデータのための索引技術, 情報処理, Vol.42, No.10, pp.953-957 (2001)
- 9) Christos Faloutsos : Searching Multimedia Databases by Content, Kluwer Academic Pub. (1998)
- 10) 並河 祐貴, 川越 恭二 : 大量のマルチメディア情報におけるインデックス手法の提案, 情報処理学会第 64 回全国大会講演論文集(3) (2002)

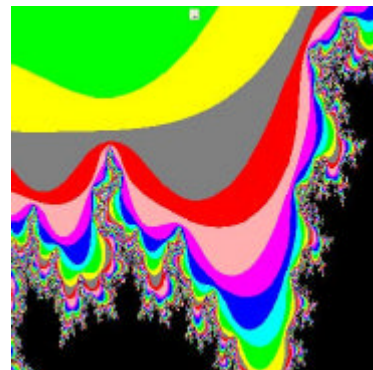
(付録) 生成したフラクタル図形の例



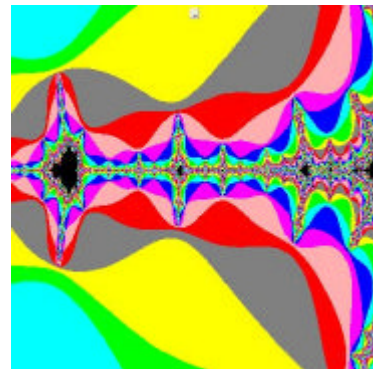
(a) クエリとしたフラクタル図形



(b) 正解集合に含まれていた類似図形



(c) 正解集合に含まれていない類似図形



(d) 非類似図形