数式画像をクエリとする類似数式検索システム

MyagmarsurenShirmenbaatar† 古賀 久志† 渡辺 俊典†

† 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
E-mail: †{shirmenbaatar,koga,watanabe}@sd.is.uec.ac.jp

あらまし インターネット上に科学技術・経済など専門分野の大量の情報が取り扱われている．数式はそのような分野のドキュメントにおいて内容を端的に示す重要な記述となっている．しかし，数式は文字列と違って分数や指数などの複雑な構造を持つため，一般の検索システムで数式をクエリとして検索することはできない．そこで，本研究では数式の画像そのものをクエリとして，類似数式画像を検索するシステムを提案する．類似数式検索に関する既存研究では数式構造をテキストとして表現するマークアップ言語を対象としたものがあるが，提案手法は数式画像を対象とする点を特色とする．提案手法では，ラベリング処理により数式画像から取り出した記号を高さ，幅，垂直位置により表現して数式間の類似度を算出する．

キーワード 類似数式，数式検索，数式画像，CBIR

Content-Based Mathematical Formula Image Retrieval System

Myagmarsuren SHIRMENBAATAR†, Hisashi KOGA†, and Toshinori WATANABE†

† Graduate School of Information Systems, University of Electro-Communications
1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: †{shirmenbaatar,koga,watanabe}@sd.is.uec.ac.jp

Abstract There are large amount of scientific and technical information on the Web. Mathematical formula is one of the significant components for representing the contents of such documents. Users can search for relevant information by using keywords and terms as queries with existing search engines such as Google and Yahoo. However, traditional search techniques are not applicable to structurally-complex mathematical formulas. In this paper, we propose a content-based mathematical formula image retrieval system. Our approach is to extract character sizes and positions as features and calculate similarity between two formulas using their feature vectors.

Key words similar formula, math search, formula image, CBIR

1. はじめに

インターネット上に科学技術・経済など専門分野の大量の情報が取り扱われている．数式はそのような分野のドキュメントにおいて，内容を端的に表す重要な記述となっている．しかし，数式は文字列と違って分数や指数などの複雑な構造を持つため，一般の検索システムで数式をクエリとして検索することはできない．そのため，ユーザの検索を支援する，内容に基づく類似検索方式が必要とされている．類似数式が検索できれば，数学問題や科学論文などの検索に有利であると考える．

このニーズに答え，数式構造をテキストとして表現するマークアップ言語を対象とした研究が進んでいる．しかし，この方式はテキスト形式のみを対象としているため，多くの Web ページで取り扱われている数式の画像データは検索の対象から外されてしまう．一方，数式は画像化した時点で数式としての情報を失い再利用できなくなる．そういった数式の画像データを MathML などのマークアップ言語に変換するためには，数式の記号や意味を再認識する必要があり，コストや手間がかかる問題がある．

そこで，本研究では数式の画像をクエリとして類似数式を検索するシステムを提案する．提案手法は記号や数式の意味を認識しなくても記号のサイズ情報と配置情報だけで類似数式が検索できる．具体的には，数式に含まれる記号のサイズと位置情報に着目した．2 つの数式からそれぞれの一番高い記号を探し，その記号の位置で合わせて比較を行う．各記号の高さ，幅，垂直位置を特徴量として抽出し，特徴ベクトルの距離を利用した数式間の類似度を算出する．より構造的に類似した数式を検索するため，類似度算出に数式内の連続した記号の高さ，幅，垂直位置から得られる 3 つの特徴ベクトルを利用した．

そして，提案システムを実装し，2 つの検索実験により性能

評価を行った．実験 1 では検索結果の再現率と適合率を調べ，実験 2 では大規模のデータに対する検索結果を人間の目で確認し，妥当に動作することを示した．

2. 関 連 研 究

2.1 マークアップ言語を対象とする類似数式検索

類似数式検索に関する既存研究では数式構造をテキストとして表現するマークアップ言語を対象としたものがある [1] [2]. MathML は数式を表現するための規格として普及し始めているマークアップ言語の一つで，XML の階層構造を使って数式の構造を記述できるため，その記述に基づく木構造の類似性を用いた類似検索が可能となっている．

この方式は，MathML 形式の数式のみを対象とするため，多くの Web ページで取り扱われている数式の画像データが検索対象から外されてしまう．数式は画像化された時点で数式としての情報を失い，再利用ができなくなる．そういった数式画像をテキスト化し MathML などに変換するためには，再認識が必要となりコストや手間がかかる．

それに対し，本研究では記号や数式の意味認識を行わず，数式画像の内容を解析することで，記号のサイズや並び順に着目した類似検索手法を提案する．画像をクエリとすることで，従来の方式で取り扱うことができなかった数式の画像データが検索可能になることが期待できる．

2.2 既存の類似画像検索システム

近年，CBIR（Content-Based Image Retrieval）と呼ばれる，内容に基づく類似画像検索が注目されており，インターネット上に多くのシステムが存在するようになった．既存の類似画像検索システムの代表的なものとして，Google Search by Image [6] や Tin-Eye [7] などが挙げられる．これらのシステムは主として色特徴や形状特徴をもとにマッチングを行っている．しかしこの方式では，色情報や形状情報の少ない数式画像は検索できない．

3. 提案システムの概要

数式をクエリとする従来の検索システムは，MathML などのマークアップ言語を検索対象としている．それに対し，提案システムは数式の画像データを検索対象としている．すなわち，画像の内容に基づいた特徴量を用いて類似画像を検索する CBIR(Content-Based Image Retrieval) 手法である．

一般の CBIR システムは，画像の色特徴や形状特徴を基に類似する画像を検索することが多いが，数式の画像には色情報や形状情報が少ないため，本システムは数式の構造を基に類似画像を検索する．システムの概要図を図 1 に示す．

図 1 のシステム概要図の 2 重線の矢印は，事前に収集した検索対象となる数式画像から特徴量データベースを作成するためのオフライン処理を示す．オフライン処理の流れは以下の通りである．

（1） Google 検索 API を用いて検索対象数式画像を収集
（2） 数式画像から連結領域となる記号を検出
（3） 記号の高さ，幅，垂直位置の情報を特徴量として抽出

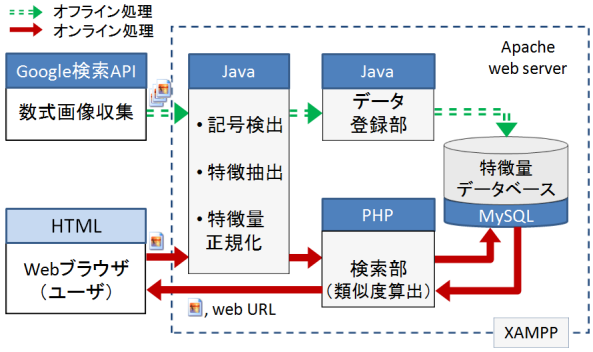


図 1 システム概要図

（4） 抽出した特徴量を正規化
（5） 正規化した特徴量を特徴量データベースに格納

一方，図 1 に示すシステム概要図の普通の矢印は，検索システムにユーザから問い合わせがあったときの特徴抽出や類似度算出などのオンライン処理を示す．オンライン処理の主な流れは以下の通りである．

（1） クエリ数式画像の提示
（2） クエリ数式画像から記号を検出
（3） オフライン処理と同じ手法で特徴量を抽出
（4） オフライン処理と同じ手法で特徴量を正規化
（5） 数式間の類似度を算出
（6） 類似度の高い順に検索結果をブラウザに表示

4. 数式の類似検索

4.1 数式の類似性の定義

類似数式を検索するために数式の類似性を定義する必要がある．人間は 2 つの数式を見比べることで，全体的にあるいは部分に似ているかどうか判断できる．しかし，数式は曖昧性を持つため類似性を言葉で表現することは非常に困難である．

提案手法は数式の意味を認識しないため，数式全体あるいは一部の「構造の類似性」に注目している．数式構造認識に関する研究 [4] では「大きい記号ほど構造を表現する役割が大きい」という数式表現において極めて基本的な特徴を利用しているが，記号は高さが幅より大きく，縦方向に記述されることが多いため，本研究では記号の高さに着目した．

まず 2 つの数式の「一番高い記号」のアスペクト比を比較し，類似していればその記号の位置で 2 つの数式を合わせて「比較領域」を決定する．比較領域が決定されたら，数式から以下の 3 つの特徴を抽出する．

（1） 記号の高さ
（2） 記号の幅
（3） 記号の垂直位置

ただし，本研究では数式に含まれるすべての文字や数字，記号などを合わせて「記号」と表記する．

そして，比較領域内の記号の高さ，幅，垂直位置で表現される特徴ベクトルごとに距離を計算し，数式間の類似度を算出する．類似度が高い数式を類似数式であると判断する．比較領域の決定について図 2 に示す数式の例を用いて説明する．数式 1

と数式 2 は，記号の外接矩形で表現されている．

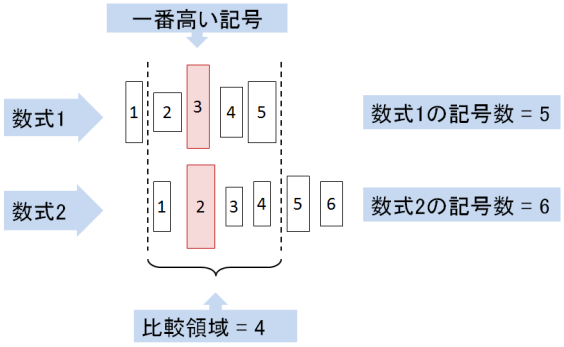


図 2　数式の比較領域

数式 1 の 3 番の矩形と数式 2 の 2 番の矩形は，それぞれの一番高い記号である．また，数式 1 の記号数は数式 2 の記号数より少ない 5 個である．2 つの数式をそれぞれの一番高い記号の位置で合わせたとき，重なる記号の数は 4 個となる．重なる記号数 4 個は，数式 1 の記号数 5 個の 8 割以上となり比較できる十分の長さであるため，点線で囲まれる部分が比較領域として決定される．

4.2　特 徴 抽 出

- ラベリングによる記号検出

ラベリングとは画像から連結領域を抽出する作業である．数式画像をラベリングした場合，同じ番号でラベル付けされる各領域が個々の記号となり，それぞれの外接矩形のサイズと位置情報がわかる．すなわち，数式は左から右へと並ぶ記号の列であるため，座標情報を利用することで記号の左上からの順番が把握できる．数式画像からの記号検出と外接矩形の座標取得の例を図 3 に示す．

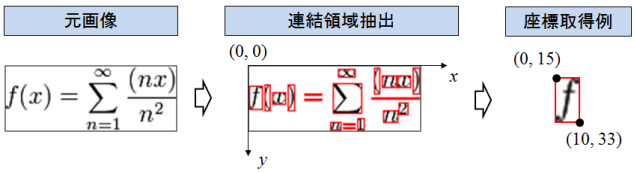


図 3　記号の検出と外接矩形の座標取得例

数式画像から記号の外接矩形の座標情報を取得する際，一番左上のピクセルを原点 (0, 0) とする．図 3 に示す数式例の一番左の「f」記号の外接矩形の左上と右下の座標はそれぞれ (0, 15) と (10, 33) である．ただし，本研究では「.」や「,」などの情報量として少ないと考えられる記号をラベリング対象外とする．また，イコール記号は 2 つの領域から構成されるが，本研究では 1 つの記号として扱う．

次に，記号の外接矩形の座標情報を用いて，高さ，幅，垂直位置を求める．記号の高さと幅をそれぞれ h, w とすれば，図 4 に示すように外接矩形の座標情報を用いて簡単に計算できる．一方，記号の垂直位置を求めるために，「Center-Line」という概念を定義した．Center-Line とは，数式の一番高い記号の外接矩形の中心点を通る横線で，Center-Line から記号の外接矩

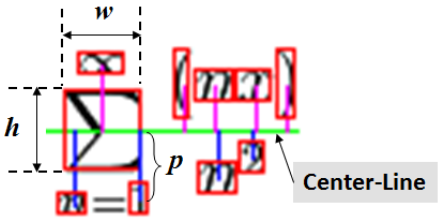
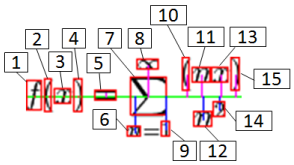


図 4　記号の高さ，幅，垂直位置

形の中心点までの距離が垂直位置 p となる．

図 4 の例では総和を表す「$\sum$」記号が一番高い記号であるため，垂直位置は 0 となる．それ以外の記号に対し，外接矩形の中心点が Center-Line より上に位置していれば垂直位置 p は正数となり，下に位置していれば負数となる．

そして，数式内の連続した記号のこれらの 3 つの情報を特徴量として抽出する．図 5 に記号の順番とその順番で抽出した特徴量の例を示す．



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
h	18	20	9	20	5	6	23	6	9	20	9	9	9	9	20
w	9	4	10	5	13	8	22	13	5	4	11	11	10	7	5
p	1	1	0	1	3	-21	0	20	-20	14	13	-14	13	-8	14

図 5　特徴抽出例

- 特徴量の正規化

数式は作成ツールによりフォントやサイズなどが異なるため，特徴量を正規化する必要がある．提案手法では，数式の一番高い記号の高さを 100 とし，それに比例するように他の記号のサイズを変えることによって，特徴量を正規化を行った．すなわち，100 と一番高い記号の元の値との比率を正規化の係数とする．数式の一番高い記号の高さを h_{max}，特徴量の正規化係数を $normCoef$ とすれば，式 (1) が成り立つ．

$$normCoef = \frac{100}{h_{max}} \tag{1}$$

記号の数が n となる数式において，高さ，幅，垂直位置特徴ベクトルをそれぞれ $H = (h_1, h_2, ..., h_n)$，$W = (w_1, w_2, ..., w_n)$，$P = (p_1, p_2, ..., p_n)$ とおき，正規化後のベクトルを $H' = (h'_1, h'_2, ..., h'_n)$，$W' = (w'_1, w'_2, ..., w'_n)$，$P' = (p'_1, p'_2, ..., p'_n)$ とおく．このとき H', W', P' の各要素は，それぞれ式 (2)，式 (3)，式 (4) によって計算する．

$$h'_i = h_i \times normCoef \quad (H \to H') \tag{2}$$

$$w'_i = w_i \times normCoef \quad (W \to W') \tag{3}$$

$$p'_i = p_i \times normCoef \quad (P \to P') \tag{4}$$

ただし，i（$i = 1, 2, ..., n$）は記号の左上からの順番を示す．図 5 の例で抽出した特徴量を正規化した結果を図 6 に示す．

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
h'	78	87	39	87	22	26	100	26	39	87	39	39	39	39	87
w'	39	17	43	22	57	35	96	57	22	17	48	48	43	30	22
p'	4	4	0	4	13	-91	0	86	-86	60	56	-60	56	-34	60

図 6　特徴量の正規化結果例

4.3　数式間の類似度

本研究で提案する数式間の類似度は，高さ特徴ベクトル，幅特徴ベクトル，垂直位置特徴ベクトルにおける異なる 3 つの距離によって算出される．高さ特徴ベクトルと幅特徴ベクトルにおける距離計算にマンハッタン距離を利用した．

クエリ数式 q と検索対象数式 t の高さ特徴ベクトル間の距離を $d(H'_q, H'_t)$，幅特徴ベクトル間の距離を $d(W'_q, W'_t)$ とする．$d(H'_q, H'_t)$ と $d(W'_q, W'_t)$ はマンハッタン距離を利用して，式 (5)，式 (6) によって計算される．

$$d(H'_q, H'_t) = (\sum_{i=1}^{m} |h'_i(q) - h'_i(t)|)/m \tag{5}$$

$$d(W'_q, W'_t) = (\sum_{i=1}^{m} |w'_i(q) - w'_i(t)|)/m \tag{6}$$

比較領域の長さ m は数式によって異なり，それによりマンハッタン距離の結果にばらつきが出てしまうため，マンハッタン距離を比較領域の長さ m で割ることで平均値を求めた．ただし，H'_q と H'_t はそれぞれクエリ数式と検索対象数式の高さ特徴ベクトルで，W'_q と W'_t はそれぞれの幅特徴ベクトルである．また，m は比較領域の長さで，i は記号の比較領域内の左上からの順番を示す．

一方，垂直位置特徴ベクトルに関しては，Center-Line より上と下でそれぞれの平均値を求め，その平均値の差で距離を表現した．Center-Line より下に位置する記号の空間位置は負数となるため，直接 2 点間の差を求める方法は適していない．そこで，クエリ数式 q と検索対象数式 t の垂直位置特徴ベクトル間の距離を $d(P'_q, P'_t)$ とすれば，式 (7) によって計算される．

$$d(P'_q, P'_t) = |Ave_{up}(q) - Ave_{up}(t)| + |Ave_{down}(q) - Ave_{down}(t)| \tag{7}$$

ただし，Ave_{up} は Center-Line より上に位置する記号の垂直位置の平均値で，Ave_{down} は Center-Line より下に位置する記号の垂直位置の平均値を示す．

垂直位置の特徴量間の距離計算例を図 7 に示す．

数式 1 と数式 2 の Center-Line はいずれも「括弧」記号の中心となるため，全ての括弧記号の垂直位置は 0 となり，計算に反映されない．したがって，数式 1 と数式 2 の Center-Line より上に位置する垂直位置の平均は，それぞれ 29 と 27 となり，下に位置する垂直位置の平均は-21 と-23 となる．そしてこれらの値をもとにマンハッタン距離を計算すると，垂直位置特徴ベクトル間の最終的な距離は $|29 - 27| + |-21 - (-23)| = 4$ となる．

そして，クエリ数式 q と検索対象数式 t 間の最終的な類似度を $Sim(q, t)$ とおけば．q と t 間の 3 つの距離 $d(H'_q, H'_t)$，

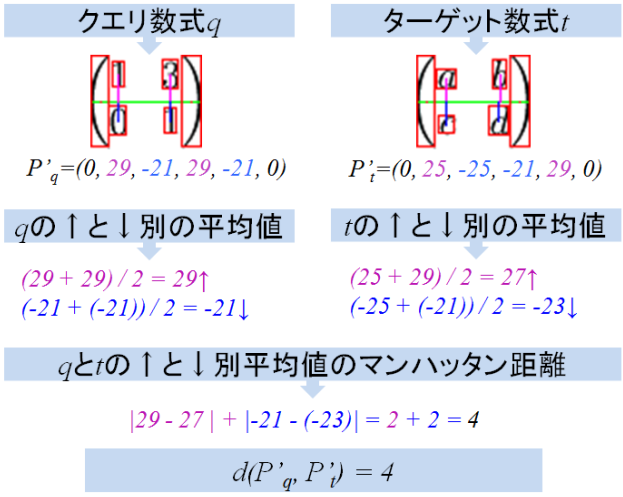


図 7　垂直位置の特徴ベクトル間の距離計算例

$d(W'_q, W'_t)$，$d(P'_q, P'_t)$ を用いて，式 (8) によって算出される．

$$Sim(q, t) = 100 - \frac{d(H'_q, H'_t) + d(W'_q, W'_t) + d(P'_q, P'_t)}{3} \tag{8}$$

この形式では，2 つの数式が同一数式である場合，3 つの特徴ベクトルにおける距離はいずれも 0 となり，類似度は 100 となる．

4.4　検索アルゴリズム

クエリ数式と特徴量データベースのデータ間の類似度算出までの処理の流れを図 8 に示す．

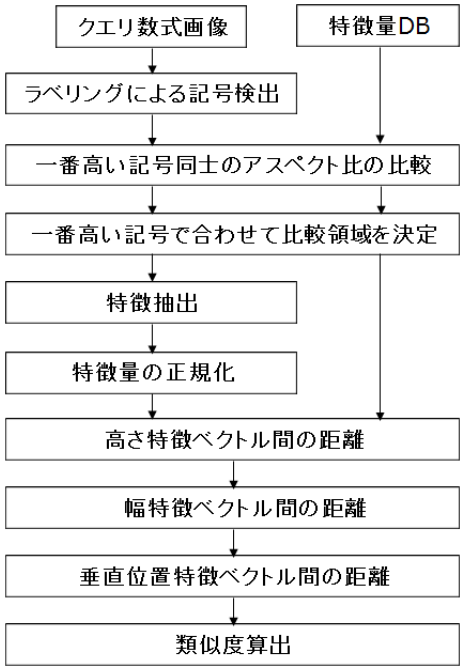


図 8　類似検索の流れ

まず，クエリ数式画像からラベリング処理によって記号を取り出す．一番高い記号のアスペクト比，特徴データベースに格納されている各数式の一番高い記号のアスペクト比と比較し，アスペクト比の差が 0.2 以下の数式において比較領域を決定す

る．そして，クエリ数式の各記号の高さ，幅，垂直位置を特徴量として抽出・正規化し，特徴ベクトル間の距離を用いて最終的な類似度を求める．

本研究で提案する数式の比較方法は，一番高い記号同士の類似を前提としているため，まず一番高い記号のアスペクト比の値を比較することで，形状として似ているかどうかを判断する．本研究では，様々な記号のアスペクト比の値を比較する予備実験を行った．記号の高さを h，幅を w とすれば，アスペクト比の値は w/h で計算される．

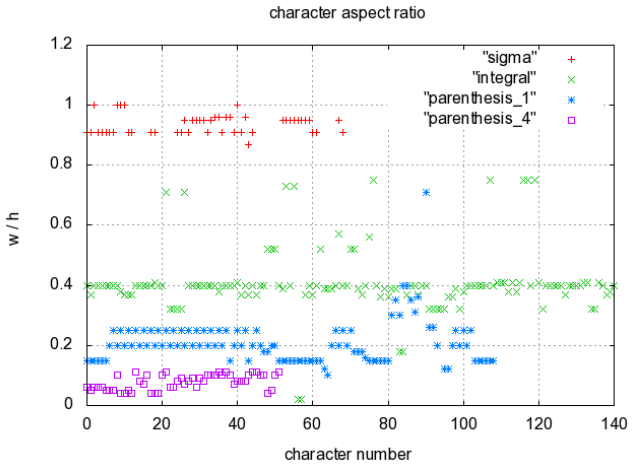


図 9　記号のアスペクト比の値

図 9 は，上から順にシグマ（69 個），インテグラル（144 個），1 行の括弧（109 個），4 行の括弧（52 個）といった 4 つの記号のアスペクト比の値の計算結果である．ただし，1 行の括弧とは，括弧に対する記号が 1 行だけになっているもので，4 行の括弧は 4 行を持つ行列の括弧などを示す．4 つの記号のアスペクト比の値は異なる領域に分布し，アスペクト比によって記号を区別可能であることがわかる．このような実験結果を基に，検索対象とする数式の一番高い記号のアスペクト比の範囲を ±0.2 と設けた．

5. 評価実験

本研究では提案システムを実装し，2 つの実験によって性能評価を行った．

- 検索実験 1

人間が正解を付与した 92 個の数式画像で構成されるデータセットを用いた定量評価実験である．

- 検索実験 2

2042 個の大規模数式データベースを用いた，人間の目による定性評価実験である．

5.1 実験方法

- 検索実験 1

検索実験 1 はクエリに対して構造的に類似する正解データを人手であらかじめ用意し，それらを対象として検索を行う実験である．実験 1 で用いる検索クエリとして，インテグラル，シグマ，括弧などの高い記号を持つ数式 6 個と，特徴的な高い記号を持たない数式 8 個の合計 14 個の数式を用いた．データベースは，クエリごとにあらかじめ用意した正解データ全 92 個の数式画像で構成される．実験で用いた 14 個のクエリ数式を図 10 に示す．クエリ 1〜クエリ 5 とクエリ 12 は特徴的な高い記号を持っているのに対し，それ以外のクエリは特徴的な高い記号を持たない．

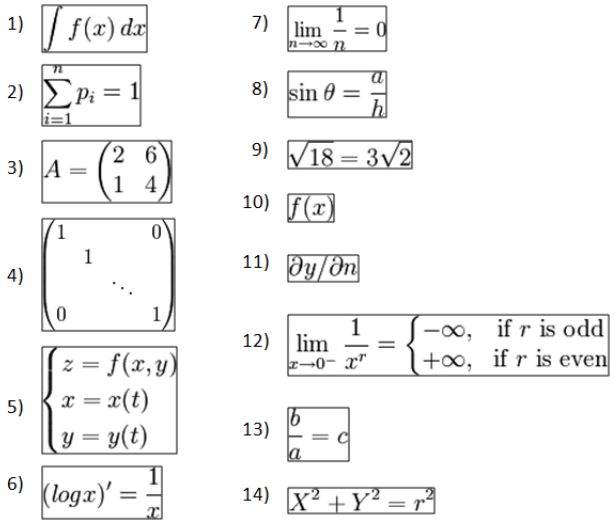


図 10　実験で用いたクエリ数式

評価方法として，類似度が 80 以上となる検索結果のなかに用意した正解数式が何個入っているかをチェックする再現率 (recall) と，正解でない結果の割合を考慮した適合率 (precision) を調べた．再現率と適合率はそれぞれ式 (9)，式 (10) によって求められる．

$$\text{再現率} = \frac{\text{検索された適合データの数}}{\text{用意した正解データの数}} \tag{9}$$

$$\text{適合率} = \frac{\text{検索された適合データの数}}{\text{検索された全データの数}} \tag{10}$$

- 検索実験 2

検索実験 2 は，あらかじめ収集した全 2042 個の数式画像を対象とした実験である．実験 2 の検索クエリは，実験 1 で用いる 14 個の数式画像となる．検索実験 2 に関しては，検索クエリに対して正解となる類似数式の数などが把握できないため，検索性能を定量的に評価することは困難である．

そのため，検索実験 2 では検索結果とクエリ数式を人間が目視で見比べることで，検索精度を定性的に評価する．なお，大規模データに対するシステムの検索時間を計測することで，提案手法のアルゴリズムとしての速度を評価する．

5.2 実験結果と考察

- 検索実験 1 の結果

14 個のクエリ数式に対して用意した正解数式の数と検索結果を利用して，再現率と適合率を計算し，それぞれの平均値を求めた．表 1 に実験 1 の結果を示す．

結果全体として，14 個の異なるクエリに対して用意したデータのほとんどが検索でき，再現率の平均は 0.76 となった．検索結果からは得に「$\int$」，「$\sum$」，「()」，「{}」などの特徴的な高い

表 1　実験 1 の検索結果

クエリ	正解数	検索結果数	適合結果数	再現率	適合率
1	7	6	6	0.86	1.00
2	12	12	12	1.00	1.00
3	9	20	8	0.89	0.40
4	9	9	7	0.78	0.78
5	4	13	4	1.00	0.30
6	4	20	4	1.00	0.20
7	6	5	3	0.50	0.60
8	9	7	7	0.78	1.00
9	5	2	2	0.40	1.00
10	6	18	5	0.83	0.28
11	3	5	2	0.67	0.40
12	5	8	5	1.00	0.62
13	8	3	3	0.37	1.00
14	5	3	3	0.60	1.00
				平均 0.76	平均 0.68

記号を持つ数式において，高い精度で類似数式を検索できていると言える．しかし，特徴的な高い記号がない数式に対する検索結果は類似しない数式を多く含んでいるため，適合率が低くなり，平均 0.68 となった．

実験 1 の結果で良かった例と良くなった例として，クエリ 1 とクエリ 6 の結果を図 11 と図 12 に示す．

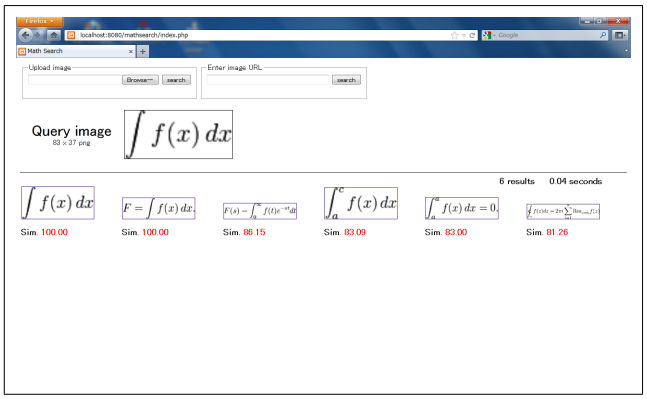


図 11　検索実験 1 のクエリ 1 の結果

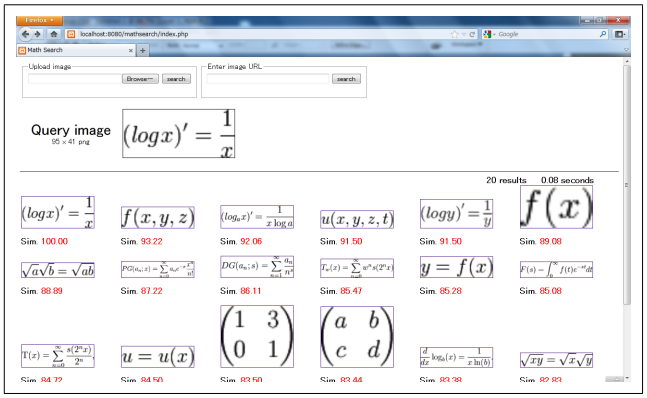


図 12　検索実験 1 のクエリ 6 の結果

クエリ 1 はインテグラル記号が特徴的な記号となっているため，用意した 7 個の正解数式のうち 6 個が検索できた．一方，クエリ 6 は特徴的な高い記号がない．用意した 4 個の正解数式（検索結果の 1, 3, 5, 17 件目）は全部検索できたが，適合率は一番低い 0.20 となった．

実験 2 においてはほとんどのクエリに対して高い精度で類似数式が検索できたが，高い記号を持たないクエリの結果は類似していない数式を多く含むことがわかった．また，検索対象データが多くなり平均検索時間は 5.71 秒となった．

6.　まとめと今後の課題

本研究では単語や専門用語などのキーワードではなく，数式画像をクエリとする類似数式検索システムを提案，実装した．検索手法として，各記号の高さ，幅，垂直位置を特徴量として抽出し，一番高い記号に着目した．一番高い記号のアスペクトが似ている場合のみ類似数式候補とし，3 つの特徴ベクトルのマンハッタン距離から類似度を求めた．

そして，類似数式検索システムを実装し，定量・定性的に性能評価を行った．インテグラルや括弧などの高い記号を持つ数式において，高い精度で類似数式を検索できることが確認できた．しかし，特徴的な高い記号がない数式における検索結果は類似していない数式画像を多く含む結果となった．適用範囲大規模のデータに対して本システムが妥当であることを示した．

今後の課題として, 以下のことが挙げられる．

(1)　高い記号がない数式における検索アルゴリズムの改善

特徴的な高い記号を持たない数式において，類似していない結果が多かったため，一番高い記号以外に有効的な特徴を見つけ出す必要がある．有効的な特徴の例として，外接矩形の形状が類似する記号同士の画像マッチングが挙げられる．

(2)　数式の構造特徴と意味情報の組み合わせ

数式の構造比較に加えて周辺単語などの意味的な情報を組み合わせることで，検索性能の向上が期待できる．

(3)　検索アルゴリズムの高速化

特徴量データベース構造を再構築し，検索対象データを絞り込む必要がある．

(4)　検索対象数式のデータ種類の拡大

数式画像だけでなく，PDF ファイルなどに含まれる数式も検索できるようにすれば，適用範囲がより広がると考える．

文　献

[1] Y. Hijikata, H. Hashimoto, and S. Nishida, “Search Mathematical Formulas by Mathematical Formulas”, Proc. of HCII 2009, LNCS vol. 5617, pp. 404-411, 2009.

[2] M. Adeel, H.S. Cheung, and S.H. Khiyal, “Math Go! Prototype of A Content Based Matematical Formula Search Engine”, JATIT, vol. 4, No.10, pp. 1002-1012, 2008.

[3] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. Kanahori, “INFTY - An integrated OCR system for mathematical documents”, Proc. of ACM, pp. 95-104, 2003.

[4] 岡本 正行，東 裕之，“記号レイアウトに着目した数式構造認識”，信学技報 J78-D-2, No.3, pp. 474-482, 1995.

[5] 渡辺 弥寿夫，中沢 政幸，“科学技術文書の画像入力における数式とフォントの認識”，信学技報, EID94-3, pp. 13-18, 1994.

[6] “Google 類似画像検索”, http://www.images.google.com/

[7] “Tin-Eye 類似画像検索”, http://www.tineye.com/