数式画像をクエリとする類似数式画像検索システムの提案

シルメンバータル　ミャグマルスレン † 坪川 宏 †
† 東京工科大学コンピュータサイエンス学部コンピュータサイエンス学科

1　はじめに

現在，インターネット上に科学技術など専門分野の大量の情報が取り扱われている．その分野に関するドキュメントにおいて，内容を端的に表す情報として専門用語と数式が挙げられる．

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，数式そのものをクエリとする，ユーザーの検索を支援する方式が必要とされている．それに答え，数式は MathML と呼ばれる XML 形式で標準化が進み，MathML 形式の数式をクエリとする検索方式が存在するようになった．しかし，現在多くの Web ページ（Wikipedia など）では数式は画像として取り扱われており，それらの画像データは検索の対象から外されてしまう．

そこで本研究では，数式の画像をクエリとして類似数式画像を検索する手法を提案し，システムを実装する．実現手法として，画像のヒストグラム間の距離を事前に計算し，記号の座標情報と組み合わせることで比較を行う．

2　既存の類似画像検索システム

近年，画像そのものをクエリとする，内容に基づく類似画像検索が注目されており，GazoPa（http://www.gazopa.com/）や　TinEye（http://www.tineye.com/）などの類似画像検索システムが存在するようになった．しかし，これらの検索システムは，色特徴と形状特徴を用いているため，色情報や形状情報の少ない数式画像は検索できない．

3　システム概要

本研究で提案するシステムの概要図を図 1 に示す．本システムは，事前に検索対象の画像データと画像の URL を収集しておく必要がある．収集された各画像において，ヒストグラムの値を算出し，ラベリング処理を行う．次に記号サイズの正規化と記号座標の平行移動変換を行い，これらの数値データを基にデータベースを作成する．

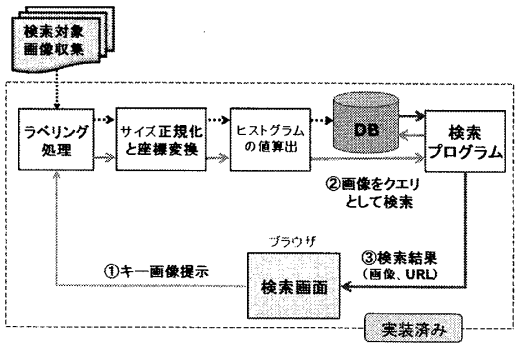

図 1: システム概要図

3.1　ラベリング処理

ラベリングとは，画像中から個別の小さな塊（blob）を抽出する作業である．本研究では，OpenCV ライブラリ [2] を用いて各 blob（記号）の左上と右下座標をもつピクセルを参照することで，その blob の包含矩形を求める．数式画像からは記号の数，面積，座標といった 3 つの情報を抽出する．

$$f(x) = \sum_{x=1}^{\infty} \frac{\lceil x \rceil}{x^2}$$

図 2: 数式画像におけるラベリングの例

3.2　サイズの正規化

インターネット上には様々なフォントやサイズの数式があり，それらのサイズを正規化する必要がある．

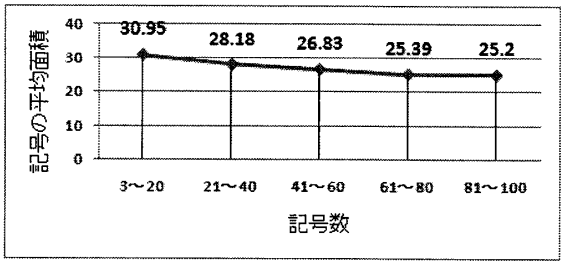

図 3: 記号平均面積の計算

3.1 節で述べたラベリング処理によって取得した記号の総面積を記号数で割った結果が平均面積となる．

A Proposal of Similarity-Based Retrieval of Mathematical Formulas using Images as Queries.
† Shirmenbaatar Myagmarsuren
† Hiroshi Tsubokawa
School of Computer Science, Tokyo University of Technology (†)

本研究では，Wikipediaから収集した全2086個の数式画像（png）を用いて画像の正規化を行う．用いた全画像の98.7%が3個〜100個の記号をもっており，これらの画像の平均記号面積を20個ごとに計算した．その結果を図3のグラフに示す．このグラフの記号平均面積の値を記号数にあわせて利用する．

3.3 座標変換

数式領域の左上のx, y座標を原点とする座標系に座標を平行移動する．これにより，画像中から数式部分だけを切り出し，不要な空白領域を取り除くことができ，記号座標を同一させることが可能となる．座標の平行移動について図4にて説明する．

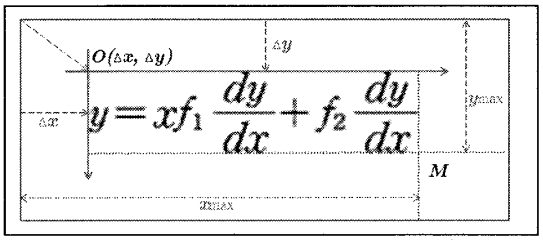

図 4: 座標変換

図4の画像では，数式領域の左上座標が$({}_{\Delta}x, {}_{\Delta}y)$となる．また，数式領域の長さと高さは，$x_{max} - {}_{\Delta}x$と$y_{max} - {}_{\Delta}y$で計算される．

3.4 ヒストグラム間の距離計算

OpenCVでは，画像のヒストグラム間の距離を数値で表すことができる．この距離を用いて4章で述べる実験1を行った．しかし，画像同士の比較に非常に時間がかかるため，各画像においてヒストグラムの値を事前に算出する必要があった．

その解決法として，各画像を白い画像と比較し，そのヒストグラム間の距離をヒストグラムの値として用いた．

3.5 画像入力

ユーザーはブラウザ上の画像アップロードホームより検索したい画像をアップロードするか，画像のURLを送信することで検索を行う．

4 検索実験

本研究で実装したシステムの動作検証として，Wikipediaから収集した2086個の数式画像を対象に2つの検索実験を行った．2つの実験を行うことで，それぞれの検索手法での結果および検索時間を比較する．以下，両実験で用いた検索手法を説明する．

- 実験1
  検索するたびに画像同士の比較を行い，ヒストグラム間の距離で検索を行う．
- 実験2
  本研究で提案したヒストグラムの値と記号の座標情報を用いて検索を行う．

5 実験結果

両実験では，記号数，サイズ，ヒストグラムの異なる10個の画像を検索クエリとして用いた．実験1の平均検索時間14.2秒に対し，実験2の平均検索時間は0.0016秒となった．また，実験1と実験2の検索結果の78 %以上が一致することが確認できた．実験2の検索結果例のスクリーンショットを図5に示す．

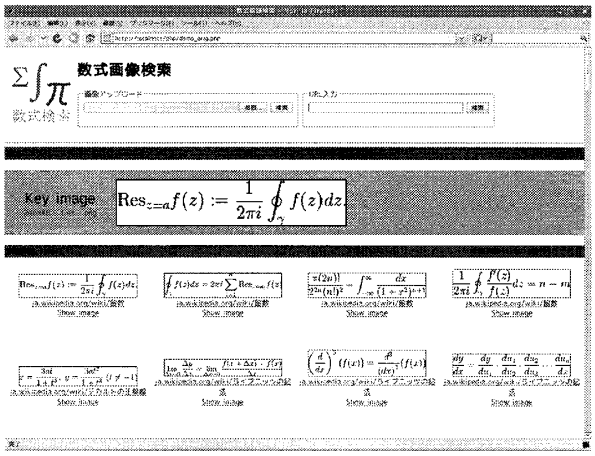

図 5: 実験2の検索結果例

6 まとめと今後の課題

数式画像をクエリとして，ヒストグラムと記号の座標情報に着目した類似数式画像検索システムを提案，実装した．検索のクエリに数式そのものを用いることでユーザーの検索を支援するシステムとなった．

しかし，記号数や形状が似ていても内容が異なる数式の場合，類似していると判断されてしまうことがある．この問題を解決するために，数式画像を1個の白い画像だけではなく，多数の画像と比較することでヒストグラムの値の算出手法を改良する必要がある．

参考文献

[1] 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）

[2] Intel Corporation：OpenCV - マルチプラットフォームな画像処理，画像認識ライブラリ，http://opencv.jp/