2I1-3

数学理解を支援するWeb上の数式画像の検索

Searching for Mathematical Expression Imagery on the Web to Support Mathematics Understanding

山田 奉子[*1] 上田 洋[*2] 村上 晴美[*3] 岡 育生[*1]
Kuniko Yamada Hiroshi Ueda Harumi Murakami Ikuo Oka

[*1] 大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University

[*2] 株式会社ATR Creative
ATR Creative Inc.

[*3] 大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University

Even though the web provides various kinds of mathematical information, searching for mathematical expressions is difficult. Since a mathematical expression cannot be replaced with words because of its complicated structure, we cannot use ordinary search systems. Our research uses an ordinary text search and presents imagery. We classify imagery based on its features that are unique to mathematical expression imagery and present suitable mathematical expression imagery related to an input keyword.

1. はじめに

数学関連の様々な情報が Web 上に存在するが，数式は検索が容易ではない．複雑な構造の数式は文字に置き換えられず，一般の検索システムは使えない．現在，pdf 形式や MathML 形式の数式については解析が進められているが，本研究はテキスト検索を用いて，Web 上に多数存在する画像形式の数式を提示する点を特色とする．数式画像特有の素性等を用いて他の画像と区別し，入力したキーワードに関連した数式画像を提示するものである．

2. 提案手法

2.1 Web 上の数式

Web 上の数式表現方法は大きく 3 つに分かれる．pdf文書，HTML 文書の中に埋め込まれた画像，XML に基づく数式記述用の MathML である．

また，一般的に数式は以下の特徴を持つ．(1) 変数等は行内で文字と同列の扱いであるが，数学ではイタリック体を用い添え字を持つものも多く，テキストではなく数式として表現される．(2) 重要な定理・公式は直前の文章が改行され新しい独立した行となり，直後も改行される．中央揃えされることも多い．(3) 定理の導出などでは長く繋がった式となる．

2.2 SVM

サポートベクターマシーン（以下 SVM）は，教師あり学習を用いた識別手法の一つであり，認識性能が優れた学習モデルとされている．今回，提示候補となる数式画像とその他の画像を区別するために，台湾国立大学の Lin らによって作られたライブラリである LIBSVM を使用する．

連絡先：山田 奉子，大阪市立大学，〒558-8585 大阪市住吉区杉本 3-3-138，d07tb553@ex.media.osaka-cu.ac.jp

2.3 概要

提案手法の概要は以下のとおりである．まず，数式で表現された定理・公式名をキーワードとしてテキスト検索をし，適切な Web サイトを見つける．次に，(1) そこから画像を取得し，その属性を得る．(2) 画像が行内か，独立行かを判定する．(3) 画像の特徴量を用いて，提示候補となる数式画像か，それ以外かを SVM で判定する．(4) 画像周辺のテキストにキーワードを含むか否かを判定する．(5) (2)(3)(4)を点数化し，順位の高いものを提示する．

(1) 画像の取得

広告を排除するために，HTML ソースに記述されている画像ファイルのみ取得し，画像のファイルサイズ，縦ピクセル数，横ピクセル数のデータも得る．また，数式 1 つが 1 つの画像とは限らないので，間にテキストがなく，タグのみで繋がれている画像ファイルは 1 つの画像とみなす．繋げた画像ファイルの縦ピクセル数は，合計ファイルサイズを横合計で割ったものとする．

(2) 画像の位置

画像が行内か，独立行かを判定する．独立行の条件は，当該画像名が記述されている<img>タグの前後が，次の 2 つを共に満たすものとし，それ以外を行内とする．(a) テキストがない．(b)
 </br>   <tr> </tr> のいずれかがある．但し center についてはタグがなくとも align 属性に記述があるものも条件に含める．

(3) SVM での判定

ここで，判別したいのは「数式らしい数式」とそれ以外である．それ以外とは以下を指す．(a) 縦横比が 1 に近くファイルサイズが小さな変数，(b) 縦横比が 1 に近くファイルサイズが大きい図・グラフ・広告・定理の導出のための一纏めの式群，(c) 縦横比が非常に大きな繋がった式群．

(a)(b)(c) を基準として学習データを用意し，「数式らしい数式」とそれ以外を区別させた．使用した素性は，「ファイルサイズ」「縦ピクセル数」「横ピクセル数」「密度」「縦横比」である．

(4) 画像周辺のテキスト

当該画像の前後各 80 字以内を限度として，キーワードを探す．但し，隣接する画像を飛び越えては探さない．また，2 つの画像の間にキーワードがある場合，最初の画像とその後の句点の間にあれば，最初の画像に，独立した文章内なら近いほうの画像に，最後の句点と次の画像の間にあれば次の画像に付与する．

(5) 提示

各画像に点数を付与し，Web ページ毎に上位の画像を提示する．点数化は，独立行か否か，SVM の出力，キーワードの有り無し，という 3 つの観点で行う．具体的には後で述べる．

3. 評価実験

3.1 データセット

大阪市立大学工学部の Web シラバスからキーワード候補を抽出し，最終的に 8 キーワードにしぼった．各キーワードにつき，5 個の Web ページを取得，計 40 個の Web ページ，1279 個の画像を得た．各画像を，人手により，キーワードに対して適切な数式かどうかを判定した．1 個の Web ページで最高 154 画像，最低 2 画像と非常にばらつきがあるため，個々の Web ページの正解数の上限は定めず文意に沿って決定した．図 1 に，例を示す．なお，今回収集した画像ファイル形式は，png，gif が圧倒的に多く，次に jpg で，1 例のみ bmp があった．画像形式による数式画像の差異は特に見られなかった．

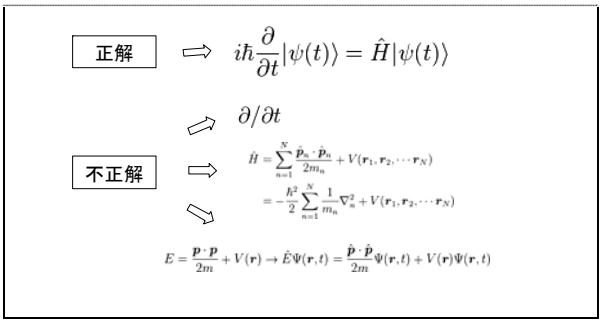

図 1: 正解・不正解例

3.2 方法と結果

各画像に付与する点数は以下の式を用いる．式中のパラメータ α_k とk = 3 の 0 or 1 は実験の結果で決定する．k = 1 は画像が独立行にあるものが正解，k = 2 はSVMの出力，k = 3 はキーワードを持つものが正解である．また，α_kは重みである．

$$points = \sum_{k=1}^{3} \alpha_k \delta_k(image)$$

$k=1,2$ の時 $\quad \delta_k(image) = \begin{cases} 1 & (image: correct) \\ -1 & (image: incorrect) \end{cases}$

$k=3$ の時 $\quad \delta_k(image) = \begin{cases} 1 & (image: correct) \\ 0 \text{ or } -1 & (image: incorrect) \end{cases}$

実験の方法は，まず各画像に上式により，点数を付与する．改行していない Web ページや，キーワードをほとんど使用していない Web ページなどがあるので，Web ページ毎に最高点が異なる．よって，それぞれの Web ページ内での最高点のグループを抽出する．それらを集めたうえで，以下の式による適合率・再現率・F 値を用いて最適なパラメータの値を決定する．ここで r は抽出した正解画像数，n は抽出した画像数，c は正解画像数とする．

$$\text{適合率} = \frac{r}{n} \qquad \text{再現率} = \frac{r}{c} \qquad F\text{値} = \frac{2\cdot\text{適合率}\cdot\text{再現率}}{\text{適合率}+\text{再現率}}$$

最初に，δ_3式で，キーワードを持たない画像に 0 か -1 のどちらを付与するかを決める．各 α_k = 1 とし 0 を付与したセットと -1 を付与したセットを作り比較した．表 1 の結果から，0 を付与した．

表 1: キーワードを持たない画像に付与する値を変えた評価

	δ_3 = 1, −1	δ_3 = 1, 0
適合率	55.3%	56.4%
再現率	71.0%	71.0%
F値	62.2%	62.9%

次に，α_kを決めるために，$\alpha_1 = \alpha_2 = \alpha_3 = 1$ の時と，各 α_kの 1 つだけを 2 に変えた時，計 4 通りで，比較した．結果を図 2 に示す．$\alpha_1 = \alpha_2 = \alpha_3 = 1$ の時のF値が一番良いので，これを採用した．

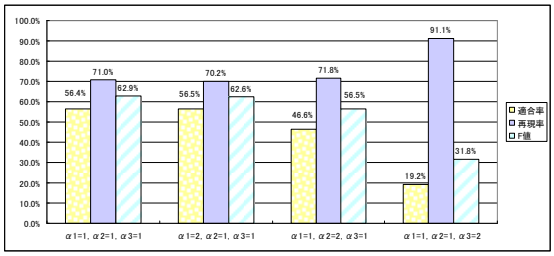

図 2: 重みを変えた場合の評価

4. 関連研究

MathML に関しては，[Yokoi 11]等の研究がなされており，国立情報学研究所による NTCIR プロジェクトの MATH タスクにも MathML 関連の論文が発表されている．しかし，多くの数学 Web 文書では，数式を画像で表現している．数式画像についての研究は少ないが，例えば[Shirmenbaatar 12]は，数式画像をクエリとする類似数式検索システムを提案し，数式画像中の一番高さが高い記号に着目している．

5. おわりに

本研究はテキスト検索を手がかりにして，キーワードに適合した数式画像を提示するものである．数式画像の文章上の位置に着目し，変数などと数式を区別した．また外形の特徴から SVM を用いて，適切な数式とその他の画像とを分離した．さらに，周辺情報を用いて，数式画像の内容を類推した．さらに，これらを組み合わせて，目的となる画像を得た．

今後の課題は，まず，キーワードの表記のゆれにより，文書中から取得できなかったキーワードが幾つかあったことである．顕著な例では，「ナビエ・ストークスの式」というキーワードに対して，1 個の Web ページの中で，4 通りの表記が混在していた．この点を改良すれば，評価値の改善が見込まれる．さらに今後，SVM 以外の他の識別手法についても，比較検討していきたい．

参考文献

[Yokoi 11] Yokoi, K., Nghiem, M-Q., Matsubayashi, Y., and Aizawa, A.: Contextual Analysis of Mathematical Expressions for Advanced Mathematical Search, in *CICLing 2011* (2011)

[Shirmenbaatar 12] Shirmenbaatar M., 古賀 久志, 渡辺 俊典: 数式画像をクエリとする類似数式検索システム，第4回データ工学と情報マネジメントに関するフォーラム，(DEIM2012) (2012)