数学用語をクエリとする Web 上の数式画像検索

Mathematical Expression Image Retrieval on Web Using Mathematical Terms As Queries

山田 奉子
Kuniko Yamada

大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University
dl8ud512@eb.osaka-cu.ac.jp

上田 洋
Hiroshi Ueda

特定非営利活動法人リンクト・オープン・データ・イニシアティブ, 株式会社 Stroly
Linked Open Data Initiative, Stroly Inc.
uedayou@linkedopendata.jp

村上 晴美
Harumi Murakami

大阪市立大学大学院工学研究科
Graduate School of Engineering, Osaka City University
harumi@osaka-cu.ac.jp

岡 育生
Ikuko Oka

(同上)
ikuko@m.ieice.org

keywords: mathematical expression image, learning support, mathematical information retrieval (MIR), web search, support vector machine (SVM)

Summary

Recently the importance of mathematical information retrieval (MIR) has been recognized and various methods for mathematical expression retrieval have also been proposed. However, since mathematical expressions on the Web are not annotated with natural language, searching for mathematical expressions by conventional search engines is difficult. For helping people in various fields who use mathematics as a learning tool, our proposed method performs a Web search using a mathematical term as a query, extracts mathematical expression images (math-images) related to the query from the obtained Web pages, and presents the top ten math-images with their surrounding information. The method measures the relevance between a query and a math-image from the following viewpoints: the math-image is in a separate line, it has the query in the neighborhood and appropriate image feature quantities, and it appears in the first part of the Web page. We use a support vector machine to discriminate if the image provides appropriate feature quantities.

We conducted two experiments with our proposed method. We determined its scoring parameters in Experiment 1 and evaluated it in Experiment 2. The results revealed the usefulness of our proposed method with accuracy, recall, F-measure, mean reciprocal rank (MRR), and mean average precision (MAP).

1. はじめに

近年, 数学情報検索 (mathematical information retrieval: MIR) の重要性が認識されてきており, Zanibbi らは「大学生が与えられた数式と同じような構造や意味を持つものを教材の中から探したいと思ったり, 研究者が与えられた関数についての論文を見つけないと願ったときに, ユーザがクエリとして与えた数式を解釈して目的のものを見つけ出すことが必要である」と述べている [Zanibbi 12]. また, 国立情報学研究所による NTCIR でも, パイロットタスクを含め過去 3 回 math タスク^{*1}が行われており, Aizawa らはその意義を「数式の検索を可能にすることは, 単に検索システムで扱える対象を拡大するというだけではなく, 数学の知識基盤構築に向けた第一歩ともなる」と述べている [Aizawa 13].

MIR について様々な手法が提案されてきたが, それらの手法は一般の Web を対象としたものではない. Web 上には数学に関する大学の講義資料や解説サイトなどの学習者にとって有益な Web ページが数多く存在し, それらの文書中の数式を画像で表現したものも多い. 数学は理工系や経済系の様々な概念を表現する道具であることから, それらの分野の学習者にとって, 数式を理解することは必須である. しかし, 数式名をクエリとして Web 検索を行っても, スニペットに数式は出てこず, 数式を見つけるために次々と Web ページを開いて探さなければならないのが現状である. そのような場合に, 検索結果を表示するスニペットとして数式を含むものの一覧が出れば役に立つと考えられる. さらに, 数式が画像であれば取り扱いも簡単であり, コンパイルの必要もなく, コンピュータの操作に習熟していなくても扱えるという利点もある.

*1 <http://ntcir-math.nii.ac.jp/>

本研究は、様々な分野の学習の道具として数学を利用する人々のために、数学用語をクエリとして Web 検索を行い、得られた Web ページからクエリに関連する数式画像を抽出し、上位 10 件を周辺情報と共に提示するものである。重要な数式は、独立した行に書かれ、数式名が近くに記述されることが多く、ほど良い大きさを持ち、文書の最初の方に書かれることが多い。本手法はクエリと数式画像の関連度を、数式が「独立行にある」、「近傍にクエリを持つ」、「適切な画像特徴量を持つ」、「文書の最初の方に出現する」という観点を用いて測るものである。

本論文の構成は以下のとおりである。2 章で関連研究について述べ、3 章で提案手法について説明する。4 章で評価実験を行う際の条件を決定し、5 章で評価を行う。さらに 6 章で議論し、最後に 7 章で全体をまとめる。

2. 関連研究

2.1 類似数式検索

数式検索とは、クエリまたは対象文書に数式を含む検索 [Aizawa 13] のことであり、狭義には、与えられた数式に対して類似構造や意味を持つものを検索すること [Zanibbi 12] である。類似数式検索を、クエリの入力形式で分類すると、以下のようになる。

§1 MathML で記述された数式

Mathematical Markup Language (MathML) について、World Wide Web Consortium の公式 Web ページ *2 には、「数式を記述しその構造と内容の両方をキャプチャするためのマークアップ言語である。HTML はウェブ上でテキストが提供、受信、処理されることを可能にしてきたが、MathML の目的はこれらの機能を数式に対して可能にすることである」と記されている。数式の構造をとらえることができるため数式検索でよく使われている。

岸本らは、MathML のタグの出現頻度から抽出されたメタデータに対して Latent Semantic Indexing (LSI) を適用することによって、データセット (W3C の MathML Test Suite 及び M. R. Spiegel 著「数学公式・数表ハンドブック」より収集) から数式を検索している [岸本 03]。Hijikata らは、あらかじめ WOLFRAM Research の MATHML CENTRAL より収集したデータセットに対して MathML の XPath で索引付けをしておき、高速にマッチングを行っている [Hijikata 09]。また横井らは、MathML の木構造から得られるパスの集合の Jaccard 係数の値で数式の類似度を測り、同時に数式周辺テキストから抽出した情報も利用してデータセット (情報処理学会論文誌より「機械学習」に関係するワードを文書中に持つ論文より構築) に対して検索を行っている [横井 11]。

§2 LaTeX で記述された数式

Zanibbi らはデータベース中の個々の LaTeX 形式の数式を bag of words に変換して個々の文書とみなし入力されたクエリと TF-IDF を用いてマッチングした結果と、それらの数式を画像に変換後 Content-Based Image Retrieval (CBIR) の手法で解析した結果の両方を用いて、クエリと arXiv より収集したデータセット内の数式との類似度を測っている [Zanibbi 11]。

§3 数式画像

Shirmenbaatar らは CBIR の手法で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、あらかじめ収集したデータセット内の数式との類似度を算出している [Shirmenbaatar 12]。Rani らは、数式画像を含む入力文書から数式画像を抽出した後、CBIR の手法で画像の特徴量を抽出し、それらの値を用いて類似するものを画像データベースから出力している [Rani 16]。

§4 条件指定

[Murata 12, 小田切 08] の手法は、MathML を拡張した問い合わせ言語を実装し、「sin を含む積分」のような数式の部分形や関数の引数をクエリとし、The Wolfram Functions Site より収集したデータセット内を検索することができる。MlaS は、MathML 形式の数式表記を含む文書を処理するフルテキストベースの検索エンジンであり、異なる表記法を持つ 2 つの数式をマッチさせるプロセスを持ち、Mathematical Retrieval Corpus (MREC) 内を検索することができる [Sojka 11]。Kim らは、平易な言葉 (plain words) をクエリとする。例えばユーザが $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ の $\sqrt{b^2 - 4ac}$ だけしか思い出せないときに、入力クエリを「root b squared minus 4ac」とし、それを MathML で処理後、高校以上の様々な数学マニュアルより収集した数式データセット内を検索するものである [Kim 12]。MathWebSearch は、LaTeX ベースの数式とキーワードを入力することにより、キーワードを含むスニペットと関連数式を the arXMLiv corpus より構築したデータセット内を検索して出力するものである [Hambasan 14]。

§1 から §4 で述べた類似数式検索についての研究は、MathML や LaTeX を入力方法としており、一般の人々にとっては扱いにくい [Pattaniyil 14]。また、検索対象についても、科学論文や専門的なデータベースなどである。これらの類似数式検索の手法には、一般の Web を対象としたものは見当たらなかった。それに対し、本手法は、数学用語をクエリとして Web 検索をし、学習の理解に役立つ数式とスニペットを提示するものである。

2.2 数式に注釈を付与

Nghiem らは、数式とその周辺テキストが同一内容を示す関係を取り出す手法を提案している [Nghiem 10]。Wikipedia の記事を入力とし、concept, description 及び数式の三つ組を用いて、テキストマッチングとパターン

*2 <https://www.w3.org/TR/MathML3/>

マッチングの手法で、名前と記述を添えた数式のリストを出力する。前述した [横井 11] では、数式とその周辺テキスト中の説明記述候補の名詞との関係の特徴量として SVM を用い、数式の定義や説明を抽出している。Kristianto らは、Springer の出版物より得られた LaTeX 文書を入力とし、事前に設定した 7 個のパターンと数式周辺テキストをマッチングさせ数式の description を抽出する方法と、すでに注釈をつけた訓練データを用いて分類器を作成する機械学習の手法で description を抽出する方法の 2 つを用いて、注釈を付与している [Kristianto 12a, Kristianto 12b]。また、description を用いて数式のセマンティック検索が可能であることも示唆している。数式にうまく注釈を付与できれば、テキスト検索で数式を検索することが可能となるが、これらの手法は構造化された文書の特長や特定のジャンルの文書特有の言い回しを利用している。しかし Web 文書は、一般的に非定型で多様性を持つので、これらの手法をそのまま適用するのは難しいと考えられる。

2.3 画像検索

画像検索の方式には、画像に付与されたテキストに基づく検索 (Text-Based Image Retrieval: TBIR) と画像の内容に基づく検索 (CBIR) がある。

TBIR で最も重要な部分が画像のラベリングであり、自動的にラベルを付与する研究が行われてきた。例えば Munson らは、「画像ファイル名」、「文書タイトル」、「画像 alt テキスト」などの HTML のメタデータを利用してラベリングを行っている [Munson 01]。しかし、alt テキストについては、付与されていれば精度は良いが、空のものが多く Web 上の画像の内容を確定するには不向きであると述べている。また相良らは、ラベルとなる名詞について「画像からの距離が近い」、「画像の関連テキスト領域中の出現頻度が高い」、「テキスト領域内の重要文に出現する」という観点で評価値を付けている [相良 04]。竹内らは、alt テキスト中の単語列と本文から抽出した単語列に対して χ^2 検定を行い確定したキーワードと、画像特徴量による類似画像検索で得たキーワードを用いて説明文を得ている。この中で alt テキストが記述されていない画像が全体の 45% あったと言及している [竹内 06]。画像のラベリングでは本研究と類似した手法が使われているが、要求されるラベリングの粒度が数式の場合とは異なる。[相良 04] では「ヒマワリ」、「ネコ」、「信濃川」、「オーロラ」、「ハト」、「フェラーリ」を識別する実験を行っているが、数式検索では、例えば、円の方程式 ($x^2 + y^2 = r^2$) を検索しているときに $x^2 - y^2 = r^2$ (双曲線) を得ても見当違いである。

CBIR については、Rani らが「CBIR は今や画像処理分野で重要な役割を果たしており、CBIR による数式検索は新しい研究分野である」と述べている [Rani 16]。CBIR を用いた類似数式画像検索の手法には、前述した [Rani

16, Shirmenbaatar 12, Zanibbi 11] がある。

3. 提案手法

3.1 数式

数学文書は、通常の文書とは異なる特徴を持つ。重要な数式は文章の途中であっても改行され、独立した行に記述される。 $\int_a^b$ や $\sum_{i=0}^n$ などの記号や、ギリシャ文字、 $\aleph$, $\mathcal{L}$, $\mathfrak{M}$ などの文字も用いられる。また、 x , y などの変数には、通常イタリック体が用いられる。

Web 上でこれらの記号や文字を表現する方法は様々であり、ソフトウェアを利用して TeX をテキストや画像に変換する場合もある。例えば、MathJax^{*3}では、公式 Web ページに、「入力として MathML, TeX, ASCIImath を用い、出力として HTML+CSS, SVG, MathML を生成することができる^{*4}」と記されている。

最終の形で数式の表現形式を分類すると 4 つに分けられる。(1) テキストで表す。(2) PDF で表す。TeX で数式を含む文書を作成後、出力された PDF を用いたり、ワードプロセッサで作成した文書の中に数式作成ソフトウェアで作成した数式を埋め込み後、PDF に変換して用いる。(3) 画像を用いる。代表的な画像形式に、PNG, JPG, GIF, SVG がある。(4) MathML で表す。MathJax で書かれたものは、数式上で右クリックすると、MathML のコードか TeX のコマンドのどちらかを選んで表示することができる。

3.2 分類器

画像が正解数式らしさを持つかどうかを分類器を用いて判別する。予備実験では、ベイズネット、ナイーブベイズ、ニューラルネットワーク、SVM (多項式カーネル)、rbf カーネルを用いた SVM、k 近傍法、決定木、ランダムフォレストを比較し、その結果より、[Chang 11] に記述された rbf カーネルを用いた Libsvm で実験を行う。10 分割クロスバリデーションでパラメータの最適値を探したが、結局デフォルト値が一番良かったため、 $C = 1.0, \gamma = 0.2$ としている。また、訓練データを構築する際に負例が非常に多く不均衡なため、負例をアンダーサンプリングしている。なお、多数の数式画像は白黒 2 値に見えるが実際にはカラー成分を含んでいるものが多く^{*5}、色情報は特徴量として用いていない。

3.3 提案手法の説明

提案手法の説明を図 1 に沿って行う。

*3 <https://www.mathjax.org/ams-becomes-managing-partner-of-the-mathjax-consortium/>

*4 <https://www.mathjax.org/#docs>

*5 予備実験のサンプルでは全体の 25%

図 1 提案手法の概要

§1 キーワード検索

キーワードをクエリとして Web 検索を行う。検索エンジンが返したスニペットを含む検索結果一覧より、各 Web ページの url、スニペット中の太字部分、html ソースファイル、画像を取得する。

§2 画像データの取得

各 Web ページのソースファイル中の img タグから画像情報を抽出し、画像名・画像特徴量（ファイルサイズ、横ピクセル数、縦ピクセル数）を取得する。ここで画像名とは img タグ中の「src=」の後のパスを除いた部分である。得られた画像名を検索エンジンによる Web ページランク順かつ当該の Web ページ内の出現順に並べた画像名リストを作成し、あらかじめ人手でキーワードに対する関連度を、正解・不正解の 2 値で付与しておく。次に以下の 3 つの要素について調べ、その結果を付与しておく。

(1) 画像の位置

画像が行内か独立行かを判定する。html ソースファイル中の img タグの前後にテキストがなく改行を示すタグがあれば独立行とする。それ以外は行内とする。改行を示すものとして次のタグを使う。

```

<blockquote>, </blockquote>, <br>, </br>, <br />,
<center>, </center>, <dd>, </dd>, <div>, </div>,
<li>, </li>, <p>, </p>, <tr>, </tr>

```

(2) 画像周辺のキーワード

当該画像の前後各 100 文字以内及びその画像が属する img タグ中の alt テキストにキーワードがあるかどうかを調べる。画像のある Web ページ内でキーワードを持つ画像が得られない場合には、代替キーワードを用いて再度

調べる。代替キーワードには、検索エンジンが返したスニペットの太字部分を用いる。この太字部分には、Web ページのどの部分がクエリとしたキーワードと関連しているかが示されており、キーワードに対する表記のゆれも含まれる。例として、キーワード「ヘッセ行列」をあげると、代替キーワードとして「Hessian 行列」、「ヘッセ」、「行列」、「Hesse 行列」、「Hesse (半角空白) 行列」、「Hessian」、「Hesse」が得られる。

(3) 正解数式らしさ

キーワードに対する正解画像候補とそれ以外の、見かけの違いを SVM で判別する。使用する特徴量は、「ファイルサイズ」、「横ピクセル数」、「縦ピクセル数」、「密度 (式 (1))」、「横縦比 (式 (2))」である。

$$\text{密度} = \frac{\text{ファイルサイズ}}{\text{横ピクセル数} \times \text{縦ピクセル数}} \quad (1)$$

$$\text{横縦比} = \frac{\text{横ピクセル数}}{\text{縦ピクセル数}} \quad (2)$$

ここで SVM を使用する目的は、正解数式らしさを持つものとそれ以外を判別することにある。図 2 及び表 1 にキーワード「erf 関数」の画像例を示す。A が正解画像であり、B 以降は不正解画像である。A は正解数式らしさを持つ。B はグラフであり、比較的ファイルサイズが大きく、密度が小さく、横縦比も 1 に近い。C は変数であり、ファイルサイズが小さく横縦比も 1 に近い。D は付加的な内容の数式で式変形もあり細長いので、横縦比が大きい。E は本文と関係が薄い画像で、ファイルサイズが大きく、縦長であり横縦比が 1 に近い。F は式変形を伴い、横縦比が 1 に近いが、ファイルサイズが B と C の間の値である。これらの特徴量の違いに着目し、正解数式らしさを持つ画像を判別する。

§3 不要画像の除去

ロゴやボタン類などのアイコンは数が多くこれらをもって除去するために、それらの画像名に共通する文字列を不要語とするリストを作成し、そのリストに従って不要画像名を画像名リストから除去する。不要語リストの作成には、SVM の訓練データを用いる。

(1) 訓練データの中で、独立行にありキーワードを持つが不正解である画像の画像名を抜き出す。(2) 括弧子及び画像名中の数字を取り去り、画像名を分解する。「- (マイナス)」、「_ (アンダースコア)」、「. (ピリオド)」で結合

表 1 図 2 の各画像データ

画像名	ファイルサイズ (バイト)	横ピクセル数	縦ピクセル数	密度	横縦比
A	1,215	201	47	0.13	4.28
B	8,779	400	280	0.08	1.42
C	206	11	9	2.08	1.22
D	3,295	655	50	0.10	13.10
E	12,440	163	197	0.39	0.83
F	2,146	288	233	0.03	1.24

(1) (2) (3): 「誤差関数 - Wikipedia」より, (4): 「誤差関数 - 高精度計算サイト」より, (5): 「恒常法の S 字カーブフィッティングについて」より

図 2 画像例 (キーワード: erf 関数)

されているもの及び「yougoSakuin」のように途中に大文字が入っているものはその部分で分解し頻度順にした上で, 出現頻度が 2 以上かつ 2 文字以上に絞り, 不要語候補リストを作成する. (3) 訓練データから正解画像名を取得し, 不要語候補リスト中の文字列とのマッチングを行う. (4) マッチした文字列を候補リストから除去し, 残った文字列のリストを不要語リストとする. (5) 画像名リストの画像名中に不要語リストの語を含むものを削除する.

129 個の不要語中, 例として頻度の上位 10 個の不要語を挙げる. 大文字・小文字は区別しない.

logo, btn, spacer, button, top, menu, search, title, sakuin, gnav

§4 点数の付与

§2 であらかじめ付与しておいた 3 つの要素について調べた結果に従い, 当該の画像が, 独立行にあれば 1 点, キーワードを持てば 1 点, 正解数式らしさがあれば 1 点を与える. すべての条件を満たせば, 3 点満点となる.

次にボーナス点について説明する. ここまでの処理で

各キーワードについて画像のランキングをすると次のようになる. 最初の並び順は, 検索エンジンが返してきた Web ページのランク順かつその Web ページ内での出現順である. 上記の 3 条件で点数を付与して降順に並べ替えると, 順位は変わるが元の検索エンジンによる上位ランクの Web ページ内の画像がやはり上位を占めることになる. その中に不正解画像が混ざっていても排除できず, 下位の正解画像が上位に上がれない. そこで, 「重要な事柄は文書内で最初の方に出現する」と仮定し, ボーナス点を導入する. これは同一の Web ページ内で最初に出現した満点の画像に, ボーナス点として追加 1 点を与えるものである. その結果, 下位にある正解画像を上位に上げることができる.

以上のことから各画像 i_k に式 (3) により得点 $score(i_k)$ を与える. ここで x_{line} は画像が独立行にあれば 1, なければ 0, x_{key} は画像の前後各 100 文字以内にキーワードがあれば 1, なければ 0, x_{svm} は SVM による分類結果が正例なら 1, 負例なら 0, x_{bo} はボーナス点があれば 1, なければ 0 とする.

$$score(i_k) = x_{line} + x_{key} + x_{svm} + x_{bo} \quad (3)$$

§5 ランキング及び提示

画像名リストの画像を, 取得した点数の降順にソートし上位 10 件を提示する. キーワード「核関数」について, 画像名リストの上位 10 件を表 2 に示す. $score$ が画像の得点であり, 降順に並んでいる. 8 位までがボーナス点を含めた満点で, 9 位・10 位はその次のボーナス点なしの満点となっており, 1 位・2 位と同じ Web ページが出現しているが, 異なる画像が選ばれている. 表 2 を用いた画面表示例を図 3 に示す. 付帯条件や式の解説などを示すために, 当該の画像の前後各 100 文字分も表示している. 表示に際しては, 文章中の変数なども画像であるので, 文意が通るように元の Web ページの `img` タグを使用しテキストと共に画像も表示している. 付随的な他の画像と区別するために, 当該画像の `img` タグを `<center>` と `</center>` で囲み `img` タグ中に `border="1" style="padding:5px;"` を書き加えることにより, 当該画像をセンタリングし余白付の枠線で囲む. また, タイト

表 2 画像名リスト (キーワード: 核関数)

	正誤	score	順位	Web ページの title
1 位	○	4	1	積分変換-Wikipedia
2 位	○	4	6	核関数とそのフーリエ級数
3 位	○	4	8	核関数に基づく方法
4 位	×	4	38	第 4 回 非線形写像とカーネル関数
5 位	×	4	39	ディリクレ核
6 位	○	4	60	積分変換 (物理のかぎしっぽ)
7 位	○	4	90	再生核ヒルベルト空間
8 位	×	4	96	rikei-index.blue.coocan.jp
9 位	○	3	1	積分変換-Wikipedia
10 位	×	3	6	核関数とそのフーリエ級数

順位: 元の Web 検索結果の順位

キーワード: 核関数

1. 積分変換 - Wikipedia

積分変換出典: フリー百科事典『ウィキペディア (Wikipedia)』移動先: 案内、検索数学の分野における積分変換 (せきぶんへんかん、英: Integral transform) とは、次の形をとるような変換T のことである:

$$(Tf)(u) = \int_{t_1}^{t_2} K(t, u) f(t) dt.$$

この積分変換の入力は関数 f であり、出力は関数 Tf である。積分変換は作用素の一種である。多くの便利な積分変換が存在する。個々の積分変換は、その変換の核関数 (kernel function) あるいは核 (kernel, nucleus) と呼ばれる

2. 核関数とそのフーリエ級数

平面波を境界値とする場合の近似解の拘束円上での誤差評価: 目次: 円外掃着波動問題の基本解近似解法による近似問題核関数 g とそのフーリエ級数正数 κ と $\gamma \in (0, 1)$ を固定して、核関数 $g(\theta)$ を以下のようにおく。

$$g(\theta) = H_0^{(1)} \left(\kappa \left| 1 - \gamma e^{-i\theta} \right| \right), \quad -\infty < \theta < \infty.$$

g の複素フーリエ係数を g_n , $n \in \mathbb{Z}$ とする。正整数 N を固定したときの g の離散フーリエ係数を

$G_n^{(N)}$, $n \in \mathbb{Z}$ とする。すなわち $G_n^{(N)} = \frac{1}{N} \sum_{j=0}^{N-1} g(\theta_j) e^{-in\theta_j}$, $n \in \mathbb{Z}$ である。これ等のフーリエ係数に対する条件 (g) と (G) を次のように置く。(g) $g_n \neq 0$, $n \in \mathbb{Z}$; (G) $G_n^{(N)} \neq 0$, $n \in \mathbb{Z}$. Graf の加法公式 ([4], Chapt. XI)

3. 核関数に基づく方法

固定して、データから K を決定する。今、領域 R として、点 $\mathbf{x}$ を中心とする辺の長さが h の超立方体 (hyper cube) を考えよう。この時、領域 R の体積は、 $V = h^M$ (29) となる。原点を中心とする辺の長さが 1 の超立方体は、核関数

$$H(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq 1/2 \quad j = 1, \dots, M \\ 0 & \text{otherwise} \end{cases}$$

(30) を用いて表すことができる。従って、 $H(\frac{\mathbf{x}-\mathbf{x}_i}{h})$ はデータ点 $\mathbf{x}_i$ が $\mathbf{x}$ を中心とする一辺 h の超立方体の内側にある時にのみ 1 となり、それ以外の場合は 0 となる。このような核関数 $H(\mathbf{u})$ は、Parzen の窓関数 (Parzen window) と呼ばれる

4. 第4回 非線形写像とカーネル関数 [levelfour_machine-learning-2014 Wiki](#) [GitHub](#)

し、 $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_M(\mathbf{x}))^T$ というように学習データを非線形変換したものを入力データとして考える。線形識別関数 $f(\mathbf{x}_i) = \mathbf{w}^T \phi(\mathbf{x}_i) + b$ を用いてこの入力データ群が線形分離可能であれば、第4回 - 線形SVMのKKT条件1より、最適な識別超平面の法線ベクトル $\mathbf{w}_0$ は

$$\mathbf{w}_0 = \sum_{i=1}^N \alpha_i t_i \phi(\mathbf{x}_i)$$

と書ける。このとき識別関数は $f(\mathbf{x}) = \mathbf{w}_0^T \phi(\mathbf{x}) + b = \sum_{i=1}^N \alpha_i t_i \phi^T(\mathbf{x}_i) \phi(\mathbf{x})$ となる。f に含まれる非線形項を次のように取り出す。 $K(\mathbf{x}_i, \mathbf{x}) = \phi^T(\mathbf{x}_i) \phi(\mathbf{x})$ この関数を核関数/カーネル関数と呼ぶ。双対問題は以下の関数の最大化になる。 $L_d(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \sum_{\alpha_i, \alpha_j \in \{0,1\}} K(\mathbf{x}_i, \mathbf{x}_j) \text{ s.t. } 0 \leq \alpha_i \leq C, \alpha^T \mathbf{t} = 0$ 線形の場合とくらべて $(\mathbf{x}^T \mathbf{T}) \mathbf{x} \rightarrow K$ となっているだ

スペースの都合上、上位 4 件を表示

図 3 画面表示例 (キーワード: 核関数)

ルをクリックすると元の Web ページを見ることができ。この例では、1 件目から 3 件目が正解画像である。4 件目は不正解画像であるが、周辺情報の中に核関数の情報が含まれている。

4. 実験 1

実験 1-1 により式 (3) の各変数の妥当性について、実験 1-2 により式 (3) の変数、 x_{line} , x_{key} , x_{svm} の重み付けの妥当性について、実験 1-3 により Web 検索で得られた上位 100 件分のデータの妥当性について検証する。

4.1 データセット

キーワードを収集するためにビショップの『パターン認識と機械学習』上下巻 [ビショップ 12a, ビショップ 12b]

(Pattern Recognition and Machine Learning: PRML) の索引を使用した。無作為にキーワード候補を選び、さらに次の 2 条件を満たすものをキーワードとした。(1) キーワード候補をクエリとした Web 検索で上位 30 件以内に数式がある。(2) 検索結果が 100 件以上ある。条件 (1) はキーワード候補が数式と関連あるとは限らず、調べる必要があったためである。検索結果一覧ページの中には PDF 文書やダウンロードしないと見られない様々な拡張子を持つファイルなどへのリンクが混在している。予備実験の結果、上位 30 件まで調べれば少なくとも 10 個の html で書かれた Web ページを閲覧できることがわかったので目安とした。条件 (2) は実験データ数を揃えるためである。以上の操作により 100 個のキーワードを得た。

次に各キーワードをクエリとして Web 検索を行い、検索エンジンが返した結果一覧のページから、各 Web ページの url, スニペット中の太字部分, html ソースファイル, 画像を取得し、人手により画像がキーワードに関連する数式画像であるかどうかの正解判定を 2 値で行った。判定は第一著者が行い、判断に迷うケースは他の 1 名 (全判定を通じて同一人物) との協議で判定した。

100 キーワードのうち、最初の 30 キーワードの上位 100 件のデータ (url, スニペット中の太字部分, html ソースファイル, 画像) *6 を用いて構築したデータセットを PRML1 (表 3), 同様に上位 101 件から最後まで用いたデータセットを PRML2 (表 4), 次の 30 キーワードの上位 100 件を用いたデータセットを PRML3 (表 4), 残り 40 キーワードの上位 100 件を用いたデータセットを PRML4 (表 4) とする。PRML1 は実験 1-1, 1-2, 1-3 で使用し, PRML2 は実験 1-3 で使用する。PRML3 は SVM の訓練データ及び不要語リスト作成に使用する。PRML4 は実験 2 で使用する。また、以下の各実験は、出力結果の上位 10 件の画像を用いて評価する。

4.2 実験 1-1

図 1(2) の 3 つの条件 (画像の位置・キーワード・正解数式らしさ) 及び (4) のボーナス点について、妥当性を検証する。

§1 方法

PRML1 を用いて、式 (3) の 4 個のパラメータのうちの 1 つだけを 0 とした 4 種類とすべてのパラメータを含むものの計 5 種類について実験を行う。a は $x_{key} + x_{svm} + x_{bo}$ で独立行の条件なし, b は $x_{line} + x_{svm} + x_{bo}$ でキーワードの条件なし, c は $x_{line} + x_{key} + x_{bo}$ で正解数式らしさの条件なし, d は $x_{line} + x_{key} + x_{svm}$ でボーナス点の条件なし, e は $x_{line} + x_{key} + x_{svm} + x_{bo}$ ですべての条件を含むものとする。実験の結果得られた上位 10 件の値を評価するために、精度 (式 (4))・再現率 (式 (5))・F

*6 以後、Web から得られたこれらをデータと呼び、これらを加工して構築したものをデータセットと呼ぶ。

表3 データセット PRML1 の内訳

キーワード	画像	正解	*	**	html
1 erf 関数	686	27	20	2	78
2 lasso	656	8	7	2	91
3 SVM	570	10	20	3	77
4 エントロピー	1,261	27	18	2	80
5 重みベクトル	432	4	40	0	60
6 カーネル密度推定法	334	12	50	3	47
7 コーシー分布	589	7	58	0	42
8 最急降下法	506	13	36	1	63
9 ジェンセンの不等式	424	9	40	11	49
10 正規方程式	721	18	51	0	49
11 相互情報量	575	21	54	2	44
12 ソフトマックス関数	463	3	38	5	57
13 凸関数	580	16	52	2	46
14 フィッシャーの線形判別	459	5	47	3	50
15 プロビット関数	399	9	37	1	62
16 ヘッセ行列	552	19	53	1	46
17 確率の乗法定理	312	14	46	1	53
18 ガウスカーネル	502	3	42	1	57
19 カルバック-ライブラーダイバージェンス	521	8	43	2	55
20 交差エントロピー誤差関数	494	14	45	3	52
21 混合分布	353	4	50	1	49
22 自己組織化マップ	230	7	50	2	48
23 潜在変数	336	3	51	0	49
24 独立同分布	243	8	57	2	41
25 汎関数微分	463	9	56	2	42
26 ボルツマン分布	574	12	45	1	54
27 マルコフ連鎖	505	12	39	1	60
28 Heaviside ステップ関数	637	21	40	1	59
29 LMS アルゴリズム	459	12	48	1	51
30 Parzen 窓	591	16	36	4	60
合計	15,427	351	1,269	60	1,671

画像: html ファイル中に記述されダウンロードできた全画像数, 正解: 全画像中で正解のもの数, *: pdf・スライド・動画・Google ブックス・ドキュメントの数, **: 各種エラーの数, html: 100 件から*および**を除いた html 形式の Web ページの数

表4 データセット PRML2, 3 及び 4

データセット	画像	正解	*	**	html
PRML2	15,724	53	940	57	1,991
PRML3	19,470	442	1,091	80	1,829
PRML4	23,988	929	1,662	72	2,266

表の項目は表3と同じである。但し, PRML2 の html の数は, 総 Web ページ数 2,988 から*および**を除いたものである。

値 (式 (6))・Mean Reciprocal Rank (MRR: 平均逆順位) (式 (7))・Mean Average Precision (MAP: 平均精度の平均)を導入する。MAP は本来, 全検索件数に対して用いる指標であるが, [酒井 15] の (2.11) 式 (式 (8)*7) で $md=10$ として各キーワードについての上位 10 件の平均精度 (Average Precision: AP) を計算し, さらにそのマクロ平均を計算することで求める。なお, 各 a, b, c, e についてはボーナス点を与える条件が変わるので, その都度再計算する。また, 以後の実験についても同様の評

価方法を用いるものとする。

$$\text{Precision}@n = \frac{r}{n} \quad (4)$$

$$\text{Recall}@n = \frac{r}{c} \quad (5)$$

$$F - \text{measure}@n = \frac{2 \cdot \text{Precision}@n \cdot \text{Recall}@n}{\text{Precision}@n + \text{Recall}@n} \quad (6)$$

$$\text{MRR} = \frac{1}{n} \sum_{k=1}^n \frac{1}{r'_k} \quad (7)$$

$$\text{AP}@md = \frac{1}{\min(md, c)} \sum_s I(s) \text{Prec}(s) \quad (8)$$

ここで, r は上位 n 件の正解画像数, c は全正解画像数, r'_k は k 番目のキーワードにおける最上位にある正解の順位, md は測定長, $I(s)$ は第 s 位の画像が正解か否かを表すフラグ, $\text{Prec}(s)$ は第 s 位の精度である。

§2 結果

表5 及び図4 が実験結果である。表中の各値はそれぞれのキーワードのマクロ平均である。F 値は, 精度の平均と再現率の平均より算出した。上位 2 件の精度・再現率・F 値は d が最も良かったが, それ以外では e が d と同等もしくは上回っていた。MRR は e が最も良く, MAP は d と e が良かった。実際の出力では, 30 キーワード中 b が 3 個, c が 1 個, d が 2 個, 上位 10 件中で正解を得られなかった。しかし, e は全キーワードについて, 少なくとも 3 位には正解を得られた。よって, 以後の実験ではすべての条件を使用して *score* を算出する。

表5 実験 1-1 の結果 (精度, 再現率, MRR, MAP)

	精度の平均				
	a	b	c	d	e
上位 1 件	0.63	0.60	0.67	0.77	0.77
上位 2 件	0.45	0.35	0.47	0.60	0.55
上位 3 件	0.40	0.30	0.38	0.50	0.53
上位 4 件	0.36	0.28	0.38	0.43	0.48
上位 5 件	0.33	0.27	0.33	0.39	0.44
上位 6 件	0.33	0.24	0.30	0.38	0.42
上位 7 件	0.31	0.24	0.26	0.37	0.40
上位 8 件	0.30	0.23	0.27	0.35	0.38
上位 9 件	0.29	0.22	0.26	0.33	0.37
上位 10 件	0.28	0.21	0.26	0.32	0.34
	再現率の平均				
	a	b	c	d	e
上位 1 件	0.11	0.10	0.09	0.13	0.13
上位 2 件	0.15	0.10	0.13	0.20	0.18
上位 3 件	0.19	0.14	0.15	0.24	0.24
上位 4 件	0.23	0.18	0.20	0.26	0.28
上位 5 件	0.26	0.20	0.23	0.30	0.31
上位 6 件	0.29	0.22	0.26	0.33	0.35
上位 7 件	0.32	0.26	0.26	0.35	0.38
上位 8 件	0.34	0.28	0.30	0.37	0.41
上位 9 件	0.35	0.30	0.32	0.40	0.43
上位 10 件	0.37	0.31	0.35	0.42	0.45
MRR	0.77	0.68	0.74	0.83	0.87
MAP	0.27	0.20	0.23	0.36	0.36

*7 他の式に合わせて変数の文字を変更している。

a: $x_{line} = 0$, b: $x_{key} = 0$, c: $x_{sum} = 0$, d: $x_{bo} = 0$, e: すべてを含む

図4 実験 1-1 の結果 (F 値)

4.3 実験 1-2

式 (3) の x_{line} , x_{key} , x_{svm} の各変数に重みを付けた場合を検討する。但し, x_{bo} については, ボーナス点を持つということは, その定義 (Web ページ内で最初に出現した満点の画像に与えられる) から $x_{line} = x_{key} = x_{svm} = 1$ を意味し, すでに最高点を取っているので検討しない。

§1 方法

各変数に重みを付けた次の式を導入する。

$$score(i_k) = a_1 x_{line} + a_2 x_{key} + a_3 x_{svm} + x_{bo} \quad (9)$$

実験時の $a_k (k = 1, 2, 3)$ の値の範囲を, 例 (表 6) を用いて考察する。 i_1 から i_4 は, 式 (9) の重みを加えても順位の変更はない。合計 1 点のものは 10 位内に出現しなかったため, 順位変更の可能性は合計 2 点のものの間のみである。ここで, i_6 が 5 位となる条件を考える。式 (9) より $score(i_5) = a_2 + a_3$, $score(i_6) = a_1 + a_3$ 。よって $score(i_6) > score(i_5)$ となるには $a_1 > a_2$ であればよい。 a_k を正の整数とすれば, 最小値で $a_1 = 2$, $a_2 = 1$ が得られる。 a_3 については特に条件がないので最小で $a_3 = 1$ とする。以上より $a_1 = 2$, $a_2 = a_3 = 1$ が得られる。他のパターンについても同様に $a_2 = 2$, $a_1 = a_3 = 1$ と $a_3 = 2$, $a_1 = a_2 = 1$ が得られる。

よって, f は $2x_{line} + x_{key} + x_{svm} + x_{bo}$, g は $x_{line} + 2x_{key} + x_{svm} + x_{bo}$, h は $x_{line} + x_{key} + 2x_{svm} + x_{bo}$ とする。これらの式を PRML1 に用いて実験を行う。

表6 上位 10 件の例 (架空のもの)

	画像名	正誤	x_{line}	x_{key}	x_{svm}	x_{bo}	合計
1 位	i_1	○	1	1	1	1	4
2 位	i_2	×	1	1	1	1	4
3 位	i_3	○	1	1	1	0	3
4 位	i_4	×	1	1	1	0	3
5 位	i_5	×	0	1	1	0	2
6 位	i_6	○	1	0	1	0	2
7 位	i_7	○	1	1	0	0	2
8 位	i_8	×	1	0	1	0	2
9 位	i_9	○	0	1	1	0	2
10 位	i_{10}	○	1	1	0	0	2

表7 実験 1-2 の結果 (精度, 再現率, MRR, MAP)

	精度の平均				再現率の平均			
	f	g	h	e	f	g	h	e
上位 1 件	0.77	0.77	0.77	0.77	0.13	0.13	0.13	0.13
上位 2 件	0.55	0.55	0.55	0.55	0.18	0.18	0.18	0.18
上位 3 件	0.53	0.53	0.53	0.53	0.24	0.24	0.24	0.24
上位 4 件	0.48	0.48	0.48	0.48	0.28	0.28	0.28	0.28
上位 5 件	0.43	0.43	0.44	0.44	0.30	0.31	0.31	0.31
上位 6 件	0.41	0.41	0.42	0.42	0.34	0.35	0.35	0.35
上位 7 件	0.39	0.39	0.40	0.40	0.37	0.38	0.38	0.38
上位 8 件	0.38	0.38	0.38	0.38	0.40	0.41	0.40	0.41
上位 9 件	0.36	0.36	0.36	0.37	0.43	0.44	0.43	0.43
上位 10 件	0.34	0.34	0.34	0.34	0.44	0.45	0.46	0.45
MRR	0.87	0.87	0.87	0.87				
MAP	0.35	0.35	0.36	0.36				

f: x_{line} の重み $a_1 = 2$, g: x_{key} の重み $a_2 = 2$, h: x_{svm} の重み $a_3 = 2$, e: すべての重みが 1, $a_1 = a_2 = a_3 = 1$

図5 実験 1-2 の結果 (F 値)

§2 結果

表 7 で, 精度は e が最も良く, 再現率は g が良かったが, いずれもその差はわずかなものであった。F 値 (図 5) は e が最も良く, MAP の値も参考にして総合的に e が良いと考え, 各重みは $a_1 = a_2 = a_3 = 1$ とし, $score$ は式 (3) に確定する。

4.4 実験 1-3

検索エンジンが返した検索結果の上位 100 件分のデータの妥当性について検証する。

§1 方法

式 (3) を用い, 上位 100 件 (PRML1) を使用した e と全件 (PRML1 及び 2) を使用した j の比較を行う。

§2 結果

表 8 及び図 6 より, すべての値について e が j と同等もしくは上回っていた。上位 10 件の正解数の合計は e が 103 個, j が 97 個であった。101 件以降のデータは表 4 中の PRML2 が示すように, 表 3 の PRML1 に比べ正解数が 20% ほどしかなく, またその分布も偏っており, PRML2 から新しい正解画像データを得られたキーワードは 30 キーワード中 4 個のみであった。101 件以降のデータを用いても結果の向上は得られないとわかったの

表 8 実験 1-3 の結果 (精度, 再現率, MRR, MAP)

	精度の平均		再現率の平均	
	j	e	j	e
上位 1 件	0.77	0.77	0.11	0.13
上位 2 件	0.55	0.55	0.15	0.18
上位 3 件	0.53	0.53	0.21	0.24
上位 4 件	0.47	0.48	0.23	0.28
上位 5 件	0.43	0.44	0.26	0.31
上位 6 件	0.41	0.42	0.31	0.35
上位 7 件	0.37	0.40	0.32	0.38
上位 8 件	0.35	0.38	0.34	0.41
上位 9 件	0.34	0.37	0.38	0.43
上位 10 件	0.32	0.34	0.38	0.45
MRR	0.87	0.87		
MAP	0.32	0.36		

j: 各キーワードについて全件使用, e: 上位 100 件使用 (実験 1-1 の e)

図 6 実験 1-3 の結果 (F 値)

で, 上位 100 件のデータを使用する.

4.5 ま と め

実験 1 の結果より, キーワードをクエリとする Web 検索の結果の上位 100 件のデータを用いて, 式 (3) により $score$ を付与すればよいことがわかった.

5. 実 験 2

実験 1 で実験の条件が確定したことを受けて, 実験 2 で提案手法の評価を行う. 使用するデータセット PRML4 は 100 キーワードのうち, 最後の 40 キーワード分の上位 100 件のデータである. 実験 2-1 では, 提案手法と他の方法との比較をし, 実験 2-2 では, 提案手法と Wikipedia との関係について考察する.

5.1 実 験 2-1

提案手法と他の 3 つの方法を比較する.

§1 方 法

m, n, p の 3 種類の方法を導入し, 提案手法を用いた k と比較する.

m は不要画像の除去を行い, 独立行の条件とキーワードの条件のみの式 (10) を用いて $score$ を付与し, 上位 10

件を取得する.

$$score(i_k) = x_{line} + x_{key} \quad (10)$$

n はキーワードをクエリとして検索を行い, 得られた結果の一覧の Web ページから手動で画像を探す. 一覧ページの上位から url をクリックして, 当該のページの最初の独立行にある数式画像について人手で正解判定をする. 数式画像がない場合は不正解とし, 次の順位のページを見る. これを繰り返し 10 件見た時点で終わりとする. 但し, 一覧ページの情報から pdf, slide, 動画, Google ブックスなど最初から数式画像を得られないとわかるページは見ずに次のページに移り, 10 件中には含めない.

p はクエリを「キーワード 数式」として Google 画像検索を行い, 上位 10 件の画像について人手で正解判定を行う.

§2 結果と考察

表 9・図 7・図 8 より結果を述べる. 図 7 中の上位 1 件の F 値は k が 0.1, m が 0.09 であった. よって, 精度・再現率・F 値・MRR・MAP のすべての値について提案手法を用いた k が最も良かった.

k と m を比較すると, 再現率については, 上位 1 件のみこの 2 つは同等であったが, それ以外では k の各値が常に m を上回っていた. さらに, 初めて正解画像が出現

表 9 実験 2-1 の結果 (精度, 再現率, MRR, MAP)

	精度の平均				再現率の平均			
	k	m	n	p	k	m	n	p
上位 1 件	0.70	0.68	0.53	0.03	0.05	0.05	0.03	0.00
上位 2 件	0.75	0.55	0.36	0.05	0.12	0.08	0.04	0.00
上位 3 件	0.66	0.51	0.29	0.05	0.16	0.11	0.05	0.01
上位 4 件	0.64	0.44	0.28	0.06	0.20	0.13	0.06	0.01
上位 5 件	0.57	0.42	0.28	0.06	0.22	0.15	0.09	0.01
上位 6 件	0.53	0.40	0.26	0.05	0.24	0.16	0.10	0.02
上位 7 件	0.52	0.40	0.25	0.05	0.27	0.21	0.11	0.02
上位 8 件	0.50	0.40	0.23	0.04	0.29	0.23	0.11	0.02
上位 9 件	0.49	0.38	0.22	0.04	0.31	0.24	0.12	0.02
上位 10 件	0.48	0.38	0.22	0.04	0.34	0.26	0.13	0.02
MRR	0.84	0.74	0.63	0.10				
MAP	0.42	0.30	0.14	0.02				

k: 提案手法, m: 独立行とキーワードの条件を使用, n: 手動で検索, p: Google 画像検索

図 7 実験 2-1 の結果 (F 値)

図 8 実験 2-1 の結果 (正解初出順位)

した順位の分布 (図 8) を見ると、上位 10 件中で、m は 40 キーワード中 3 個正解を得られなかったが、k はすべてのキーワードについて正解を得ることができた。これらの結果から、提案手法の特徴である SVM による「正解数式らしさ」の判定とボーナス点の付与が有効に働いていることが示された。

比較方法 n に関しては、上位 1 件については精度が k の 75.7%，MRR が k の 75.0% の値が得られたが、それ以外は余り良い結果を得られず、また 6 個のキーワードについて上位 10 件中で正解を得られなかった。文字情報を検索するなら、検索結果一覧ページのスニペットを見て効率良く目的の情報を得ることができると考えられるが、数式画像を検索する場合はスニペットに数式が出てこないで、あまり効率的でないことが示された。

比較方法 p に関しては、数式よりも説明図やグラフが多く、29 個のキーワードについて上位 10 件中で正解を得られなかった。ユーザが画像検索したいときによく使われる方法であるが、数式画像を検索する場合には効率的でないことが示された。

5.2 実験 2-2

百科事典サイトとして Wikipedia を取り上げる。Web 検索の結果で上位に出現することも多く、「Wikipedia ページだけを見れば良いのではないか」という疑問がある。そこで提案手法と Wikipedia を比較するために 2 つの実験を行う。

§1 方法

実験 2-2-1 は、PRML4 に提案手法を用いた k、PRML4 から Wikipedia ページを除いたものに提案手法を用いた q、PRML4 の Wikipedia ページのみに提案手法を用いた r、Wikipedia ページ内の独立行にある画像を出現順に並べた s を比較する。

実験 2-2-2 は、実験 2-2-1 と同じ手法で行うが、データセットは、PRML4 の中で Wikipedia ページが正解画像を持つキーワードのデータのみに限定する。実験 2-2-1 に準じた各手法を k', q', r', s' とし 10 個のキーワード

のデータを PRML4 から除去したもので実験を行う。その結果、30 キーワード中 22 個において、元の検索結果の 1 位を Wikipedia が占めるデータセットとなった。

§2 結果と考察

実験 2-2-1 (表 10・図 9・図 10) については、上位 1 件の F 値は k が 0.09、q が 0.11 であったので、上位 1 件では q がすべての値で最も良く、上位 2 件以降は k が F 値・MAP で最も良かった。上位 1 件に限っては Wikipedia ページを利用しない方が良い結果が得られた。r, s の差はあまりなかったが、上位 2 件以降の F 値については、「全 Web ページ使用 > Wikipedia ページを除いた Web ページ使用 > Wikipedia ページのみ使用」という順位であった。また、図 10 を見ると、k, q では全キーワードに対して上位 4 件までに正解を出しているが、r, s については 10 件中に正解を得られないものも多かった。

実験 2-2-2 (表 11・図 11・図 12) については、上位 1 件の F 値は k', q' が 0.10 で r', s' が 0.09 であったが、概ね実験 2-2-1 と同じような傾向が見られた。また、図 12 より、k', q' は上位 3 件までに正解を出しているが、r', s' はそうではなかった。

実験 2-2-1 において、k の上位 10 件の全正解数のうち、16.8% は Wikipedia ページのものであるが、上位 1 件では Wikipedia を利用しない方が良く、上位 10 件ではすべて

表 10 実験 2-2-1 の結果 (精度, 再現率, MRR, MAP)

	精度の平均				再現率の平均			
	k	q	r	s	k	q	r	s
上位 1 件	0.70	0.78	0.55	0.53	0.05	0.07	0.04	0.04
上位 2 件	0.75	0.70	0.44	0.50	0.12	0.14	0.05	0.06
上位 3 件	0.66	0.63	0.38	0.45	0.16	0.17	0.07	0.07
上位 4 件	0.64	0.57	0.33	0.37	0.20	0.20	0.07	0.08
上位 5 件	0.57	0.52	0.29	0.32	0.22	0.19	0.08	0.08
上位 6 件	0.53	0.50	0.26	0.28	0.24	0.22	0.08	0.08
上位 7 件	0.52	0.49	0.24	0.25	0.27	0.25	0.09	0.09
上位 8 件	0.50	0.49	0.23	0.23	0.29	0.27	0.10	0.09
上位 9 件	0.49	0.47	0.22	0.21	0.31	0.29	0.10	0.10
上位 10 件	0.48	0.46	0.21	0.21	0.34	0.31	0.10	0.10
MRR	0.84	0.87	0.64	0.63				
MAP	0.42	0.38	0.15	0.17				

k: 提案手法, q: Wikipedia 以外提案手法, r: Wikipedia のみ提案手法, s: Wikipedia 独立行出現順

図 9 実験 2-2-1 の結果 (F 値)

図 10 実験 2-2-1 の結果 (正解初出順位)

図 12 実験 2-2-2 の結果 (正解初出順位)

表 11 実験 2-2-2 の結果 (精度, 再現率, MRR, MAP)

	精度の平均				再現率の平均			
	k'	q'	r'	s'	k'	q'	r'	s'
上位 1 件	0.77	0.83	0.73	0.70	0.05	0.05	0.05	0.05
上位 2 件	0.78	0.70	0.58	0.67	0.11	0.10	0.07	0.08
上位 3 件	0.70	0.67	0.51	0.62	0.14	0.12	0.09	0.10
上位 4 件	0.69	0.58	0.43	0.51	0.18	0.13	0.10	0.10
上位 5 件	0.61	0.53	0.38	0.43	0.19	0.15	0.10	0.11
上位 6 件	0.57	0.53	0.35	0.38	0.20	0.18	0.11	0.12
上位 7 件	0.55	0.50	0.32	0.35	0.23	0.20	0.12	0.12
上位 8 件	0.52	0.50	0.30	0.32	0.24	0.22	0.13	0.13
上位 9 件	0.51	0.48	0.29	0.29	0.26	0.23	0.13	0.13
上位 10 件	0.51	0.47	0.27	0.28	0.29	0.25	0.14	0.14
MRR	0.88	0.90	0.82	0.80				
MAP	0.44	0.39	0.21	0.23				

k' : 提案手法, q' : Wikipedia 以外提案手法, r' : Wikipedia のみ提案手法, s' : Wikipedia 独立行出現順

のデータを使用した方が良かった。さらに実験 2-2-2 では, 30 キーワード中 22 個において, Wikipedia が元の検索結果の 1 位を占めていたが, 上位 1 件に限っても提案手法を上回ることはなかった。

5.3 ま と め

実験 2-1 で, すべての評価値 (精度・再現率・F 値・MRR・MAP) において, 提案手法を用いた k が最も良かった。比較方法では, 上位 10 件中に正解が得られなかったキーワードがあったが, k はすべてのキーワードについて正解が得られた。実験 2-2 では, 「Wikipedia ページ

図 11 実験 2-2-2 の結果 (F 値)

だけを見れば良いのではないか」という疑問に対して, 総合的に上位 10 件のデータで見れば, すべての Web ページを用いる方が良かったことがわかった。

6. 議 論

本研究は, 数学用語をクエリとして Web 検索をし, 得られた Web ページからクエリに関連する数式画像を抽出し, 上位 10 件を周辺情報と共に提示するものである。

本研究の特徴は, 自然言語である数学用語を用いて Web ページから数式画像を得ることができる点にある。

提案手法は, 名前のある数式であれば, 出力結果は意味的に類似数式の集合となる。与えられた数式が何であるかわからない場合には数式自体をクエリにする必要があるが, 名前を持つ数式で異なる表現を知りたい場合には本手法が有効である。類似数式検索では, 手法によっては変数として使われた文字が違えば検索できない場合があるが, 本手法では問題とならない。また, 式変形による違いについても問題とならない。例えば, キーワード「共分散」の上位 10 件の出力結果では, 図 13 のように異なる 6 個の正解数式を得ることができた。しかし, 一方で「三平方の定理」のように複数の名前を持つ数式を統一して出力することや, 「リウヴィルの定理」のように同じ名前で複数の定理が存在する場合に, 分類して出力することには対応していない。

提案手法を用いた k について, 上位 10 件中の正解数

(1) $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$	(2) $S_{xy} = \frac{1}{n}(x - \bar{x})(y - \bar{y})$
(3) $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))^T]$	(4) (標本) 共分散 $S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
(5) $\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$	(6) $\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$

(1): 「共分散とは - Wikipedia」より, (2): 「分散共分散行列 - 大人になってからの再学習」より, (3): 「cov - 共分散行列 - Scilab Online Help」より, (4): 「2 変量への応用 統計学入門」より, (5): 「共分散・共分散行列 - 99.9%の科学〜曖昧から確信へ - 統計学」より, (6): 「固有ベクトル、主成分分析、共分散、エントロピー入門」より

図 13 上位 10 件の正解画像の例: キーワード「共分散」

図 14 k の上位 10 件中の正解数式を持つ Web ページについての元の検索順位分布

式を持つ Web ページの元の検索順位（検索エンジンが返した順位）の分布を調べ、次の結果を得た（図 14）．正解数式を持つ Web ページ数の 1 位から 10 位での割合は 43.8%，11 位から 20 位では 28.1%，21 位から 100 位では 28.1%であった．このことから、元の検索結果の 11 位以降からも多くの正解数式を得られたことがわかった．

本研究の「重要な数式は、独立した行に書かれ、数式名が近くに記述されることが多く、ほど良い大きさを持ち、文書の最初の方に書かれることが多い」という仮説を、正解数式画像がどの程度満たしていたかを検証する．仮説を「正解数式が独立行にあり、キーワードを持ち、SVM が正例である」と読み替える．「文書の最初の方」を表す「ボーナス点」については独立条件でないので省く．PRML1, 4 の全正解数式について、95.3%が独立行にあり、87.5%がキーワードを持ち、67.0%が正例であった．実験 2-1 の k（上位 10 件）では、それぞれ 100%，99.9%，99.5%であった，全正解数での SVM が正例である割合が少し低いが、これは分類器の性能によるものも含まれているので、この仮説はまずまず妥当であったと考えられる．

alt テキストに関しては、PRML4 中の全画像について、alt テキストよりキーワードを得て k の上位 10 件に含まれたものがあったか調べたが、そのような画像はなかった．PRML4 中の alt テキストの 95%は空、tex の数式表記、または画像 id ナンバーなどの利用できないものであり、残り 5%が自然言語表記であったが、適切な数式名は得られなかった．

実験 1-2 では、重みを正の整数に限ったが、連続値を用いるなら Learning to Rank のような機械学習の手法が考えられるので、今後検討したい．

k の上位 10 件の画像 400 個の内訳を調べた結果（図 15）、図・グラフ・表は良く判別されていたが、11.7%のバナーが排除されずに残っており、そのうち 10.8%は SVM による判別が正例であった．また、不正解数式 34.5%のうち 18.5%は、正解とするには至らないが、キーワードに対して有益な情報を持つ数式であった．有益な情報を

図 15 k の上位 10 件で得られた 400 画像の内訳

持たない 16.0%の不正解数式の排除については、これらの外見が正解数式と変わらないものがほとんどなので、SVM の判別性能を上げることで対処することは難しい．正解数式を増やして不正解画像を減らす方法には、クエリの工夫が挙げられる．例えば、キーワード「RBF」では、100 件中 21 件が無関係なものであった．キーワードがアルファベット略語の場合、多義性により無関係なサイトを多く含んでしまうので、クエリに「工学」などの言葉を加える工夫が考えられる．

実装に関しては、画像をダウンロードする時間がネックとなるので、例えば、教科書のある授業のサポートシステムとして、あらかじめ教科書の索引を用いてシステムを構築しておき学生に提供することが考えられる．他の方法としては次の 2 点があげられる．(1) 検索件数を減らす．(2) html ソースファイルのみをダウンロードし、img タグ中の横ピクセルと縦ピクセルの値を用いて仮の特徴量として正解候補を選び、その後それらの画像のみダウンロードし処理する．(1), (2) については、実行時間と出力結果の精度の兼ね合いについて、今後の実験が必要である．

さらに今後の主な課題は、SVM の限界を克服するために Deep Learning などの機械学習の手法を用いて数式画像を判別し出力結果の向上を目指すことや、他のソースを用いてデータセットを構築し提案手法の汎用性を示すことであると考えられる．

7. お わ り に

様々な分野の学習の道具として数学を利用する人々のために、数学用語をクエリとして Web 検索を行い、得られた Web ページからクエリに関連する画像を抽出し、上位 10 件を周辺情報と共に提示する手法を提案した．提案手法の特徴は、MathML や LaTeX を用いず自然言語である数学用語をクエリとして Web 検索を行い、数式画像を出力することである．

評価実験の結果、提案手法の有効性を確認することができた．また、通常あまり読まれることのない検索結果の下位にある Web ページからもクエリに関連する数式画像を得られたことが確認できた．

今後の課題は、Deep Learning などの機械学習の手法を用いて出力結果の向上を目指すことや、他のソースを用いてデータセットを構築し提案手法の汎用性を示すことである．

◇ 参 考 文 献 ◇

- [Aizawa 13] Aizawa, A., Kohlhase, M., and Ounis, I.: 数式検索タスク NTCIR-11 Math-2, 第 5 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (SIG-AM), 05-11, pp. 58-60 (2013)
- [ビショップ 12a] ビショップ, C. M. 著, シュプリンガー・ジャパン株式会社編, 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇監訳: パターン認識と機械学習 上, pp. 334-341, 丸善出版株式会社 (2012)
- [ビショップ 12b] ビショップ, C. M. 著, シュプリンガー・ジャパン株式会社編, 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇監訳: パターン認識と機械学習 下, pp. 408-420, 丸善出版株式会社 (2012)
- [Chang 11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2, 27:1-27:27 (2011)
- [Hambasan 14] Hambasan, R., Kohlhase, M., and Prodescu, C.: Mathwebsearch at NTCIR-11, in *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pp. 114-119 (2014)
- [Hijikata 09] Hijikata, Y., Hashimoto, H., and Nishida, S.: Search mathematical formulas by mathematical formulas, in *Proceedings of 13th International Conference on Human-Computer Interaction (HCI International 2009)*, pp. 404-411 (2009)
- [Kim 12] Kim, S., Yang, S., and Ko, Y.: Mathematical equation retrieval using plain words as a query, in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pp. 2407-2410 (2012)
- [岸本 03] 岸本 貞弥, 中西 崇文, 櫻井 鉄也, 北川 高嗣, 栃木 敏子: MathML を用いた類似数式検索方式の実現, 第 14 回データ工学ワークショップ論文集 (DEWS2003), 6-P-07 (2003)
- [Kristianto 12a] Kristianto, G. Y., Nghiem, M.-Q., Matsubayashi, Y., and Aizawa, A.: Extracting definitions of mathematical expressions in scientific papers, in *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 3P1-IOS-2a3 (2012)
- [Kristianto 12b] Kristianto, G. Y., Topić, G., Nghiem, M.-Q., and Aizawa, A.: Annotating scientific papers for mathematical formula search, in *Proceedings of the 5th Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2012) of The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pp.17-18 (2012)
- [Munson 01] Munson, E. V. and Tsybalenko, Y.: To search for images on the web, look at the text, then look at the images, in *Proceedings 1st International Workshop on Web Document Analysis*, pp.39-42 (2001)
- [Murata 12] Murata, T. and Otagiri, K.: Searching and ranking mathematical formulas, in *Mathematical Information Retrieval Workshop (MIR2012)* (2012)
- [Nghiem 10] Nghiem, M. Q., Yokoi, K., Matsubayashi, Y., and Aizawa, A.: Mining coreference relations between formulas and text using Wikipedia, in *Proceedings of the 2nd International Workshop on NLP Challenges in the Information Explosion Era (NLPIC 2010)*, pp.69-74 (2010)
- [小田切 08] 小田切 健一, 村田 剛志: MathML を用いた数式検索, 第 22 回人工知能学会全国大会, 1F1-3 (2008)
- [Pattaniyil 14] Pattaniyil, N. and Zanibbi, R.: Combining TF-IDF text retrieval with an inverted index over symbol pairs in math expressions: The tangent math search engine at NTCIR 2014, in *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)* (2014)
- [Rani 16] Rani, S. J. and Kumari, V. V.: An effective mechanism of feature based retrieval of mathematical expression from documents, *International Journal of Applied Engineering Research*, Vol. 11, No. 5, pp. 3462-3468 (2016)
- [相良 04] 相良 直樹, 砂山 渡, 谷内田 正彦: HTML テキストの重要文を用いた画像ラベリング手法, 電子情報通信学会論文誌 D, Vol. J87-D1, No.2, pp.145-153 (2004)
- [酒井 15] 酒井 哲也: 情報アクセス評価方法論—検索エンジンの進歩のために—, p. 29, コロナ社 (2015)
- [Shirnenbaatar 12] Shirnenbaatar, 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), E6-2 (2012)
- [Sojka 11] Sojka, P. and Lřska, M.: The art of mathematics retrieval, in *Proceedings of the 2011 ACM Symposium on Document Engineering*, pp. 57-60 (2011)
- [竹内 06] 竹内 謹治, 黄瀬 浩一: 類似画像とキーワードを利用した Web 画像の説明文抽出, 情報処理学会研究報告, NL-171, pp.7-12 (2006)
- [横井 11] 横井 啓介, Nghiem, M.-Q., 松林 優一郎, 相澤 彰子: 意味と構造を考慮した数式検索手法の提案, 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011), B2-2 (2011)
- [Zanibbi 11] Zanibbi, R. and Yuan, B.: Keyword and image-based retrieval for mathematical expressions, in *Proceedings of Document Recognition and Retrieval XVIII*, vol. 7874 of *Proc. SPIE*, pp. 011-019 (2011)
- [Zanibbi 12] Zanibbi, R. and Blostein, D.: Recognition and retrieval of mathematical expressions, *International Journal on Document Analysis and Recognition*, Vol. 15, No. 4, pp. 331-357 (2012)

[担当委員: 西村 悟史]

2017 年 9 月 6 日 受理

—— 著 者 紹 介 ——

山田 奉子 (学生会員)

2007 年大阪市立大学大学院創造都市研究科修士課程修了。現在, 同研究科博士 (後期) 課程在学中。

上田 洋 (正会員)

2010 年大阪市立大学大学院工学研究科後期博士課程修了。博士 (工学)。同年, 株式会社国際電気通信基礎技術研究所 (ATR) 入社。株式会社 ATR-Promotions, ATR 知能ロボティクス研究所, 株式会社 ATR Creative を経て, 現在, 株式会社 Stroly 勤務。2014 年より特定非営利活動法人リンク・オープン・データ・イニシアティブ理事。

村上 晴美 (正会員)

1986 年京都大学文学部哲学科心理学専攻卒業。同年, 富士通株式会社入社。1994 年英国 UMIST 計算機学科修士課程修了。1998 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。同年, 大阪市立大学学術情報総合センター講師, 2001 年助教授, 2003 年大学院創造都市研究科助教授, 2007 年教授。2018 年より大学院工学研究科教授となり現在に至る。博士 (工学)。情報検索の研究に従事。ACM, 情報処理学会, 日本図書館情報学会などの会員。

岡 育生

1978 年大阪大学工学部通信工学科卒業。1983 年同大学工学研究科博士課程修了。同年, 電気通信大学助手, 1991 年同大学助教授, 1991 年大阪市立大学工学部助教授。1988 年~1989 年文部省在外研究員 (UCLA), 2000 年~2001 年ハワイ大学在外研究員。2004 年大阪市立大学大学院工学研究科教授。博士 (工学)。無線通信, 光通信における符号化, 変調に関する研究に従事。IEEE, 電子情報通信学会などの会員。