The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

数学用語をクエリとするWeb上のPDF文書を対象とした数式検索

Mathematical Expression Retrieval in PDF documents from Web using Mathematical Terms as Queries

山田 奉子　　村上 晴美
Kuniko Yamada　　Harumi Murakami

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

Since mathematical expressions on the web are not annotated with natural language, searching for expressions by conventional search engines is difficult. Our proposed method performs web searches using a mathematical term as a query and extracts expressions related to it from the obtained PDF documents. We convert PDF to TeX, create images from the mathematical descriptions in TeX and obtain image feature quantities. The expressions related to the query are discriminated by SVM using the feature quantities. Our experimental results showed that MRR's performance is the best when using both PDF and HTML.

1. はじめに

Web 上の数式を，通常の検索システムで効率よく検索することは難しい．以前，数学用語をクエリとしてキーワード検索を行い，得られた HTML 形式の Web ページから数式画像を抽出し，画像周辺情報と共に上位 10 件の数式画像を提示する研究を行い，一定の成果を得られた[Yamada 17]．その後，新しいデータセットを用いて再度実験を行った．今回，その際に得られた PDF 文書を用いて，HTML との比較実験を行う．PDF 文書には画像がないため，PDF を TeX に変換後，数式記述部分から画像を作成することによって行う．クエリと数式との関連度は，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」という観点より測る．

2. 提案手法

図 1 に沿って説明する．

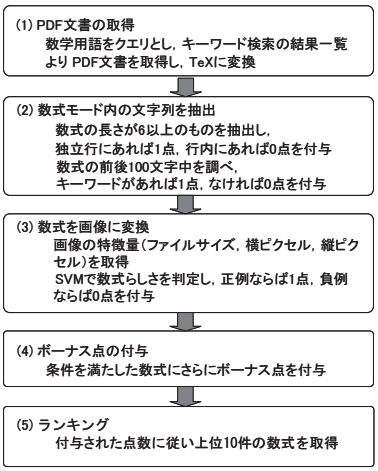

図 1：提案手法の概要

2.1 PDF 文書の取得

数学用語をクエリとしてキーワード検索を行い，上位 100 件の結果一覧から PDF 文書を取得し，InftyReader (http://www.sciaccess.net/jp/InftyReader/) を用いて TeX に変換する．

連絡先：山田 奉子，大阪市立大学創造都市研究科，
大阪市住吉区杉本 3-3-138，k16uz90z10@st.osaka-cu.ac.jp

2.2 数式モード内の文字列を取得

変数などの数式の断片を除くために，数式モード内の文字列の長さを定義する．TeX のコマンドの中で，関数の一部として用いるものや文字（例えば，\frac, \partial, \alpha, x）は 1 点，それ以外のもの（例えば，\mathrm \overline）は 0 点として，合計点を出し長さとする．$y = f(x), E = mc^2$（TeX の表記では E=mc^2）などが長さ 6 となるので，6 以上を数式として抽出する．ディスプレイ数式モードなら独立行として 1 点，インライン数式モードなら行内として 0 点を付与する．

抽出した数式の前後 100 文字中にキーワードがあれば 1 点，なければ 0 点を付与するが，キーワードを持つ数式が得られない場合は，代替キーワードを用いて再度調べ点数を付与する．ここで，代替キーワードとはキーワード検索の結果一覧のスニペット中の太文字をさす．

2.3 数式を画像に変換

数式を python のライブラリを用いて png 形式の画像に変換し，その特徴量を用いて SVM で数式らしさを判定する．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．

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

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

Libsvm を用いて判定し，結果が正例なら 1 点，負例なら 0 点を付与する．

2.4 ボーナス点の付与

この時点で各画像は検索エンジンのランク順かつ PDF 文書の出現順で並んでいる．点数を付与して降順に並び替えても，元々下位にあったものはなかなか上位に上がれない．そこで「重要な事柄は文書内の最初の方に出現する」として，ボーナス点を導入する．これは同一 PDF 内で最初に出現した満点 3 点の数式に，追加 1 点を与えるものである．これにより下位にある正解数式を上位に上げることができる．以上より，各数式 i_k に式 (3) により得点 $score(i_k)$ を与える．

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

ここで，x_{line} は独立行か否か，x_{key} はキーワードを持つか否か，x_{svm} は SVM の判定が正例か負例か，x_{bo} はボーナス点を持つか否かである．

2.5 ランキング

score(i_k) に従ってランキングをし，上位 10 件を取得する．

3. 実験

3.1 データセット

表 1:データセット

	PDF			HTML		
	正解数	全画像数	PDF 数	正解数	全画像数	HTML 数
合計	398	29,708	297	181	6,030	554
正解割合	1.34%			3.00%		

ビショップの『パターン認識と機械学習』上下巻の索引から無作為にキーワードを選びクエリとして Web 検索を行い，各キーワードについて上位 100 件の Web ページを得た．先の研究で 30 キーワード分の HTML 中の画像データを用いた分類器を作成しているので，今回これを使用する．別に 10 キーワード分の PDF 文書をテストデータとした．キーワードは「正定値行列」「多峰性」「等式制約」「ニュートン-ラフソン法」「平均二乗平方根誤差」「ラグランジュ乗数」「確率の加法定理」「カーネルトリック」「一様サンプリング」「運動エネルギー」である．得られた数式の正解判定は人手により 2 値で行った．表 1 の PDF 数は TeX に変換できた数であり，HTML 文書に比べると 1 文書が多くの画像を持つが，その中で正解の占める割合が少ないことがわかる．

3.2 結果と考察

精度（式(4)）・再現率（式(5)）・F 値（式(6)）・平均逆順位（MRR）（式(7)）・平均精度の平均（MAP）を用いて，出力結果の上位 10 件を評価する．MAP は平均精度（AP）（式(8)）を計算し，さらにそのマクロ平均を求めることで得る．

$$\text{Precision@}n = \frac{r}{n} \tag{4}$$

$$\text{Recall@n} = \frac{r}{c} \tag{5}$$

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

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

$$\text{AP@n} = \frac{1}{\min(n,c)}\sum_{s} I(s)\text{Prec}(s) \tag{8}$$

ここで，r は上位 n 件の正解数式数，c は全正解数式数，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，$I(s)$ は第 s 位の数式が正解か否かを表すフラグ，Prec(s)は第 s 位の精度である．

表 2:実験の結果（精度・再現率）

	精度			再現率		
	PDF	HTML	MIX	PDF	HTML	MIX
上位 1 件	0.20	0.50	**0. 60**	0.02	**0. 06**	**0. 06**
上位 2 件	0.40	**0. 75**	0.65	0.08	**0. 20**	0.13
上位 3 件	0.33	**0. 63**	0.60	0.10	**0. 24**	0.18
上位 4 件	0.28	**0. 63**	0.58	0.11	**0. 32**	0.23
上位 5 件	0.30	**0. 54**	0.48	0.15	**0. 37**	0.24
上位 6 件	0.28	**0. 47**	**0. 47**	0.17	**0. 38**	0.28
上位 7 件	0.31	0.44	**0. 46**	0.22	**0. 41**	0.32
上位 8 件	0.30	**0. 43**	**0. 43**	0.24	**0. 44**	0.34
上位 9 件	0.33	**0. 42**	**0. 42**	0.30	**0. 49**	0.38
上位 10 件	0.33	0.40	**0. 41**	0.33	**0. 52**	0.41

表 3:実験の結果（MRR・MAP）

	PDF	HTML	MIX
MRR	0.50	0.75	**0. 77**
MAP	0.17	**0. 34**	0.28

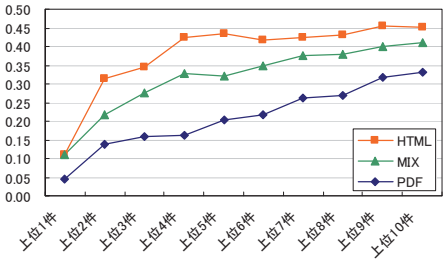

図 2:実験の結果（F 値）

表 2・表 3・図 2 中の，PDF は PDF 文書から得られた数式，HTML は HTML 文書から得られたもの，MIX はこれらを元の検索結果の順位に戻して改めてランキングしたものである．概ね，HTML の方が結果が良かったが，精度の一部及び MRR では，両方を用いた方が良かった．PDF 文書中 22.6%はスライドを PDF にしたものであり精度が悪く，上位 10 件で正解は 23 個中 2 個しか得られなかった．今後の課題は，スライド由来の PDF を排除する方策を考えること，新しい分類器で精度を上げること，長さ以外の前処理を行うことなどである．今後，テストデータの件数を増やし，新しい機械学習の方法を試してみることも必要と考える．

4. 関連研究

数式検索とは，通常類似数式検索を指すことが多いが，多くの検索システムがキーワードと数式の両方もしくは片方で検索可能である．しかしその検索範囲は特定のデータベース内である．例えば，MathWebseaerch は the arXMLiv corpus より構築したデータセット内を検索している [Hambasan 14]．また，LaTeX 形式の数式及びそれらを画像に変換したものを用いて類似数式検索を行っているものに [Zanibbi 11] の手法がある．この手法も arXiv よりデータセットを構築しており，Web 検索の結果得られたデータよりデータセットを構築しているものは見当たらなかった．

5. おわりに

数学用語をクエリとしてキーワード検索を行い，得られた PDF 文書を TeX に変換後，その数式部分の記述から数式画像を作成しその特徴量を利用することによって，キーワードに関連する数式を抽出することができた．HTML 文書を用いたものと比較した結果，MRR については両方を用いた方が良いことがわかった．

PDF は HTML より多くの正解数式を持っているので，上で述べた今後の課題について取り組み，精度を上げて行きたい．

参考文献

[Hambassn 14] Hambassa, R. et al.: Mathwebsearch at NTCIR-11, in *NTCIR-11*, pp. 114-119 (2014)

[Yamada 17] Yamada, K. et al.: Presenting Mathematical Expression Images on Web to Support Mathematics Understanding, in *IEA/AIE 2017*, Springer, pp40-46 (2017)

[Zanibbi 11] Zanibbi, R. and Yuan, B.: Keyword and Image-Based Retrieval for Mathematical Expressions, in *DRR XVIII*, pp. 011-019 (2011)