1F5-1

文字列パターンとMathMLによる構造を利用した数学問題文の検索

A Method for Searching Question Sentences in Mathematics by String Patterns and the Structure of MathML

狩山 和亮[*1] 吉仲 亮[*1] 山本 章博[*1]
Kazuaki KARIYAMA Ryo YOSHINAKA Akihiro YAMAMOTO

[*1]京都大学大学院 情報学研究科
Graduate School of Informatics, Kyoto University

We aim at developing a method for information retrieval from a set of Japanese sentences representing questions in examinations in mathematics of National Center Test for University Admissions. Our objectives are a method of analysing the structure of question sentences, and a distance between two subquestion sentences which can have mathematical formulae. We use patterns, which is a finite string of characters and variables, to comprehend common descriptions included in question sentences, and to analyse the structure of given question sentences. We define similarity between two subquestion sentences by adopting the Earth Mover's Distance as the distance between two sets of formulae. We applied our methods to texts of mathematical questions distributed by National Institute of Informatics for extracting a set of subquestions and searching for subquestions.

1. はじめに

高校や大学などにおける学校教育において，試験問題の解法を学習する際には，過去に出題された問題から類題を収集して重点的に取り組むことにより，効果的な学習を行っている．また，教員など試験を実施する側の人が試験問題を作成する際，過去に類題が出題されていないかを効率的に検査することが求められる．本研究は，日本語で記述される数学の問題文を対象とした検索手法を開発することを目標とする．本稿では，大学入試センター試験における数学の試験問題を対象にする．

数学の試験問題は，大問を単位としてまとめられた複数の小問文から構成されている．我々は，問題から類題を探すとき，大問の構造が多少異なっていても，小問文がクエリと類似していれば，その問題を学習に有用な類題とみなす．したがって，効果的な類題検索を行うためには，大問同士の比較ではなく，ひとつの大問を構成する小問文の集合を抽出した上で，小問文を単位として比較を行うことが有効であると考えられる．本研究では，小問文に相当する文をクエリとして，大問単位で与えられた問題データから入力文と類似する小問文を検索する手法を提案する．

クエリとの比較対象となる小問文の集合を抽出するために，大問中に頻出する記述を文字列パターンを用いた分類器によって抽出し，それらの記述の位置および行頭の問題番号を参照して，大問を構造化する．次に，構造化した大問から抽出した小問文集合を検索対象として，小問文間の類似度を算出する．数学の問題では数式を用いて多くの内容を表現するため，各文書に含まれる数式間の類似度を考慮した尺度を用いることで，より文書の特徴を反映した検索が行えると考えられる．本稿では，自然言語文と数式が混在する小問文を，自然言語文のみからなる文書と数式集合に分割し，それぞれに対して類似度を定義した上で二つの類似度を統合することで，小問文間の類似度を定義する．提案手法を用いた問題文検索システムの概要を図1に示す．

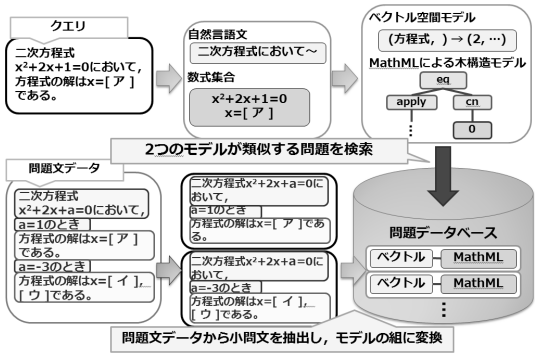

図 1: 検索システムの概要

2. 対象データ

本稿では，問題文検索に用いるデータとして，国立情報学研究所[*1]が作成した「大学入試センター試験問題用アノテーション済みデータ数学 ver2」（以下問題データ）及び株式会社ジェイシー教育研究所[*2]が販売する「大学入試センター試験問題データベース　センター Ten2011　通常版全教科セット」を利用する．問題データは，過去に実施された大学入試センター試験の数学の試験問題からなる XML データであり，問題番号と本文との位置関係や，本文中に含まれる数式と空欄の位置関係を表すようにタグ付けがされている．

大学入試センター試験の数学の試験問題は，いくつかの大問から構成されている．大問は一般的に図 2 に示すような構造をしている．大問は問題番号の後に，変数の定義や解く対象となる数式など問題を解くための前提となる記述があり，その後に続いて解答が求められる質問を表現する記述が与えられている．本稿では，前者の記述を**前提記述** (premise discription)，後者の記述を**質問記述** (questioning discription) と定義し，データ中の大問文に対して人手によるタグ付けを行う．質問記述は小問文の形式で与えられており，いくつかの小問文のリストが

連絡先: 〒 606-8501 京都市左京区吉田本町総合研究 7 号館 327 号室山本研究室 kariyama@iip.ist.i.kyoto-u.ac.jp

∗1 http://21robot.org
∗2 http://www.jc-edu.co.jp/products

中間としてまとめられていることもある．また，ひとつの質問記述と，質問記述に解答するために必要な前提記述との組を**小問文** (subquestion sentence) と定義する． 大問文に含まれる

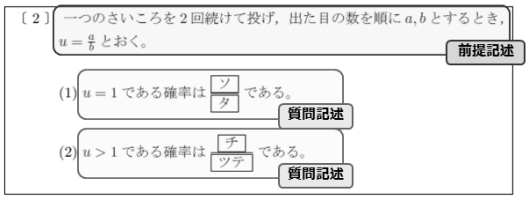

図 2: 数学の問題文に含まれる前提記述と質問記述

問題番号，前提記述，質問記述，およびそれらの位置関係を参照することで，図 2 の問題文からは,「一つのさいころを 2 回続けて投げ，出た目の数を順に a,b とするとき，$u=\frac{a}{b}$ とおく。$u=1$ である確率は $\frac{\text{ソ}}{\text{タ}}$ である。」「一つのさいころを 2 回続けて投げ，出た目の数を順に a,b とするとき，$u=\frac{a}{b}$ とおく。$u>1$ である確率は $\frac{\text{チ}}{\text{ツテ}}$ である。」という 2 つの小問文を抽出することができる．

また，問題データには，数式の一部の情報が欠落している．本稿では，小問文間の類似度に数式が持つ構造情報を用いるために，データに対して MathML の Content Markup 記法を用いて数式の埋め込みを行う．

3. 小問文集合の抽出

大問を構成する問題番号，前提記述，質問記述が判明しているとき，それらをノードとみなして大問文を木構造で表現した上で，大問文を構成する小問文集合を抽出する．

大問文中に含まれる質問記述のノードをそれぞれ葉ノードとする．ある記述のノードの親は，その記述の直前に存在する記号または記述のノードとする．例外として，小問番号のノードの親は，その小問を含む中問（なければ大問）の最後に存在する問題番号または記述のノードとする．また，中問番号のノードの親は，その中問を含む大問の最後に存在する問題番号または記述のノードとする．以上の操作を繰り返し，木全体の根である大問番号のノードまでパスを接続することによって，ひとつの大問を表現する木構造を定義する．

根ノードから葉ノードまでの経路上に存在する記述を組み合わせた文をひとつの小問文として，大問文を構成する小問文集合を抽出する．

3.1 前提記述と質問記述の抽出

前提記述と質問記述が判明していない大問文から小問文を抽出するために，既知の問題データから前提記述と質問記述を人手で抽出し，教師データとして用いる．テストデータとして用いる記述の有限集合 U を定義し，U に含まれる前提記述の有限集合を U_P，質問記述の有限集合を U_Q とする．教師データとして用いる前提記述の有限集合を T_P，質問記述の有限集合を T_Q とする．ここで，$U_P \cap U_Q = \emptyset$，$T_P \cap T_Q = \emptyset$ とする．また，記述を教師データとして用いる際，記述から数式とすべてのタグを除外する．教師データ T_P, T_Q を基に，未知の文字列 d を前提記述のクラス，質問記述のクラス，その他のクラスに分類する多クラス分類問題を解くことを目的として，パターン言語 [1] を用いた教師あり学習による分類器を作成する．

表 1: 前提記述を特徴付けるパターンの候補

前提記述 1	初項 7，公比 2 の等比数列を（数式）とする。
前提記述 2	点 C における二つの円の共通接線と直線 DE との交点を F とし，直線 DA と直線 EC の交点を G とする。
出力パターン	x_1，x_2 の x_3 を x_4 とする。

3.1.1 パターン言語

文字の有限集合を Σ で表す．また，変数の可算無限集合を X で表す．$(\Sigma \cup X)^+$ の要素を**パターン** (pattern) と呼ぶ．本研究では，同じ変数が高々1 回しか現れないパターンである**正規パターン** (regular pattern) のみを扱い，特に断らない限りこれを単にパターンと呼ぶ．文字をそれ自身に写す，半群としてのパターンからパターンへの準同形写像を**代入** (substitution) と呼ぶ．θ による p の像を $p\theta$ で表す．$p = q\theta$ を満たす代入 θ が存在するとき，q は p の**汎化** (generalization) である，あるいは q は p を**包摂する** (subsume) と呼び，$p \preceq q$ で表す．特に，$p \preceq q$ かつ $q \npreceq p$ であるとき，$p \prec q$ で表す．パターン $p_1, \ldots, p_n$ に対して，$p_1 = q\theta_1, \ldots, p_n = q\theta_n$ を満たす代入 $\theta_1, \ldots, \theta_n$ が存在するとき，q は $p_1, \ldots, p_n$ の**共通汎化** (common generalization) であると呼ぶ．さらに，$p_1, \ldots, p_n$ の共通汎化 q が $p_1, \ldots, p_n$ の任意の共通汎化 q' に対して $q \not\succ q'$ を満たすとき，q は $p_1, \ldots, p_n$ の**極小共通汎化**（minimal common generalization）であると呼ぶ．

パターン p に空代入を含む任意の代入を行うことで得られる文字列の無限集合を p で生成される**パターン言語** (pattern language) と呼び，$L(p)$ で表す，パターン集合 $\Pi = \{p_1, \ldots, p_n\}$ に対して，$L(\Pi) = L(p_1) \cup \ldots \cup L(p_n)$ と定義する．

本研究では，T_P, T_Q をそれぞれパターン集合 Π_P, Π_Q が生成する言語 $L(\Pi_P), L(\Pi_Q)$ の部分集合であると仮定し，Π_P, Π_Q を学習することで，未知の記述に対する分類器を作成する．作成する分類器は，T_P と T_Q を入力として，$|U_P \triangle (U \cap L(\Pi_P))|$ と $|U_Q \triangle (U \cap L(\Pi_Q))|$ を共にできるだけ小さくするパターン集合 Π_P と Π_Q を出力する．ここで，$S_1 \triangle S_2$ は集合 S_1 と S_2 の**対称差** (symmetric difference) を表す．

3.2 分類器の作成

本研究では，記述の集合を特徴付けるパターンの候補を決定するために，Needleman-Wunsch アルゴリズム [2] を改良したアルゴリズムを用いることで，教師データに含まれる二つの正例から，それらの極小共通汎化となるパターンを出力する．二つの前提記述からパターンを出力した例を表 1 に示す．(数式）は，記述から数式が除かれた箇所を示す．

T_P と T_Q を入力として Π_P あるいは Π_Q を出力する ExtractPattern アルゴリズムを Algorithm 1 に示す．Π_P を求める場合には，正例集合 E^+ を T_P，負例集合 E^- を T_Q とする．同様に，Π_Q を求める場合には，正例集合 E^+ を T_Q，負例集合 E^- を T_P とする．

得られたパターン集合を用いて，未知の文字列 d を前提記述のクラス，質問記述のクラス，その他のクラスに分類する．d を文字のみのパターンとした上で，Π_P と Π_Q の各要素が包摂する正例と負例の数を用いてスコア $score_P$ と $score_Q$ を定義する．

$$score_P = \sum_{p \in \Pi_P | p \succeq d} (|L(p) \cap T_P| - |L(p) \cap T_Q|)$$

$$score_Q = \sum_{p \in \Pi_Q | p \succeq d} (|L(p) \cap T_P| - |L(p) \cap T_Q|)$$

Algorithm 1 ExtractPattern(k, m, E^+, E^-)

$\Pi := \emptyset$
for $i := 0$ to k **do**
  e_1 と e_2 を E^+ から無作為に選択した文字列とする．
  p を e_1 と e_2 の極小共通汎化であるパターンとする．
  if $\frac{|L(p)\cap E^+|}{|L(p)\cap E^+|+|L(p)\cap E^-|} \geq m$ **then**
    $\Pi := \Pi \cup \{p\}$
    $E^+ := E^+ \setminus L(p)$
  end if
end for
return Π

記述 d に対して $score_P > score_Q$ ならば前提記述のクラス，$score_P < score_Q$ ならば質問記述のクラス，$score_P = score_Q$ ならばその他のクラスに分類する．

4. 小問文間の類似度

問題データから抽出した小問文間の類似度を定義する．各小問文からすべての数式を抽出し，小問文中の自然言語文間の類似度と，抽出した数式集合間の類似度をそれぞれ定義した後，二つの類似度を用いて，自然言語文と数式が混在する小問文間の類似度を定義する．

まず，形態素解析を用いて各小問文から索引語を抽出し，各索引語に対して重みを割り当てる．自然言語文に関しては自立語のみを索引語として抽出し，数式に関してはひとつの数式ごとにひとつの索引語とみなす．小問文 i に含まれる索引語 j の重み w_j^i を，tf-idf を用いて以下のように定義する．tf_j^i は小問文 i における索引語 j の出現頻度，N は全小問文数，df_j は索引語 j が含まれる小問文数とする．

$$w_j^i = \mathrm{tf}_j^i \log \frac{N}{\mathrm{df}_j}$$

小問文中の自然言語文間の類似度に関しては，索引語の重みを用いてベクトル空間モデルを作成し，コサイン類似度を尺度に用いる．文書 i の特徴ベクトル $\mathbf{D}_i = (w_1^i, w_2^i, \ldots, w_m^i)$ を用いて，文書 d_1, d_2 間の類似度を以下のように定義する．m は全語彙数とする．

$$SIM_N(d_1, d_2) = \frac{\sum_{j=1}^m w_j^{d_1} w_j^{d_2}}{\sqrt{\sum_{j=1}^m (w_j^{d_1})^2}\sqrt{\sum_{j=1}^m (w_j^{d_2})^2}}$$

小問文中の数式集合間の類似度を求めるために，数式 i, j 間の類似度 $T\text{-}sim(i, j)$ を定義する．MathML による数式の構造を用いた類似数式検索については様々な手法が研究されているが [3, 4, 5]，本稿では横井ら [3] の手法を用いて $T\text{-}sim(i, j)$ を定義する．横井らは数式を MathML の Content Markup 記法による木構造モデルに変換し，市川ら [6] の手法を用いて求めた木構造間の類似度を，数式間の類似度として提案している．

$T\text{-}sim(i, j)$ を用いて，数式集合 Π_1, Π_2 間の類似度を **Earth Mover's Distance**（以下 EMD）[7] により定義する．EMD とは，二つの分布間の類似度計算を輸送問題として捉え，最適な輸送コストを分布間の距離とする手法である．要素数 m の数式集合 $\Pi_1 = \{e_1, \ldots, e_m\}$，要素数 n の数式集合 $\Pi_2 = \{f_1, \ldots, f_n\}$ とし，w_{e_i} を索引語としての数式 e_i の重みとする．数式 e_i を輸送問題における供給地，数式 f_j を需要地としたときの e_i から f_j への輸送量を F_{ij} とし，以下の目的関数を最大化する全輸送量 $\mathbf{F} = (F_{ij})$ を求める．

$$\mathrm{WORK}(\Pi_1, \Pi_2, \mathbf{F}) = \sum_{i=1}^m \sum_{j=1}^n T\text{-}sim(e_i, f_j) F_{ij}$$

制約条件は以下のように与えられる．

$$F_{ij} \geq 0 \ (1 \leq i \leq m, 1 \leq j \leq n)$$
$$\sum_{j=1}^n F_{ij} \leq w_{e_i} \ (1 \leq i \leq m)$$
$$\sum_{i=1}^m F_{ij} \leq w_{f_j} \ (1 \leq j \leq n)$$
$$\sum_{i=1}^m \sum_{j=1}^n F_{ij} = \min(\sum_{i=1}^m w_{e_i}, \sum_{j=1}^n w_{f_j})$$

制約条件下で求めた最適な全輸送量 $\mathbf{F}^* = (F_{ij}^*)$ を用いて，数式集合 Π_1, Π_2 間の類似度 $SIM_E(\Pi_1, \Pi_2)$ を定義する．

$$SIM_E(\Pi_1, \Pi_2) = \frac{\sum_{i=1}^m \sum_{j=1}^n T\text{-}sim(e_i, f_j) F_{ij}^*}{\sum_{i=1}^m \sum_{j=1}^n F_{ij}^*}$$

例外として，$\Pi_1 = \emptyset$ かつ $\Pi_2 = \emptyset$ のとき，$SIM_E(\Pi_1, \Pi_2) = 1$ とする．また，$\Pi_1 = \emptyset$ かつ $\Pi_2 \neq \emptyset$，$\Pi_1 \neq \emptyset$ かつ $\Pi_2 = \emptyset$ のいずれかであるとき，$SIM_E(\Pi_1, \Pi_2) = 0$ とする．

最後に，**小問文間の類似度** SIM_Q を，SIM_N と SIM_E を用いて定義する．

$$SIM_Q = \sqrt{SIM_N \cdot SIM_E}$$

5. 実験

5.1 前提記述と質問記述の抽出

1997 年〜2010 年の問題データに含まれる 269 の大問文を用いて，前提記述と質問記述の抽出実験を行った．テストデータとして，大問文から人手で抽出した 1,513 個の前提記述と 1,454 個の質問記述を含む 8,242 個の記述を用いた．テストデータから無作為に選択した各 50〜1,000 個（50 刻み）の正例と負例を教師データとし，教師データの選択から識別精度の測定までを 1 シークエンスと数える．前提記述に関して，テストデータに含まれる前提記述の集合 U_P と，分類器が前提記述と判定したテストデータの集合 U_P' を用いて，以下に定義する適合率と再現率の調和平均（F 値）を識別精度を測るための尺度とした．

$$precision = \frac{|U_P \cap U_P'|}{|U_P'|}$$
$$recall = \frac{|U_P \cap U_P'|}{|U_P|}$$

質問記述に対しても同様の手法により F 値を計算し，各教師データ数に対して，100 回のシークエンスを実行することで得られた F 値の平均を識別精度とした．また，ExtractPattern アルゴリズムにおける入力は，$k = 1{,}000, m = 0.9$ とした．教師データ数を 100 としたときのパターン集合 Π_P, Π_Q の例と，Π_P, Π_Q を用いて識別を行ったときの F 値を表 2 に示す．実験の結果，前提記述に関しては約 60.1〜79.7%，質問記述に関しては約 77.3〜83.6%の精度でそれぞれの記述を識別可能であることが分かった．

表 2: パターン集合の例

パターン集合	要素	F 値
Π_P	x_1 と x_2 する x_3	0.659
	x_1 点 x_2 と x_3 を x_4。	
	x_1 い x_2 と x_3	
	の x_1，x_2	
	x_1 点で交わ x_2 る x_3	
	を x_1する x_2	
	x_1 平面上に x_2	
	硬貨を 6 回投げるとき，	
	x_1 とき x_2，	
	x_1 らな x_2 た x_3 て x_4 る。	
	x_1 ると，	
Π_Q	x_1 が x_2 り x_3。	0.821
	x_1 である。	
	このとき，となる。	
	x_1，x_2 る x_3 と x_4 が x_5。	
	x_1 となる。	
	であり，	

5.2 小問文間の類似度

2004 年〜2010 年の問題データから抽出した 564 の小問文を検索データベースとして，小問文をクエリとする類似検索システムを実装した．索引語抽出のための形態素解析は，Mecab[*3]を用いて行った．また，問題データに数式を埋め込む際，EzMath[*4]を用いて数式から MathML への変換を行った．

2010 年の問題データに含まれる 72 の小問文をクエリとしてそれぞれ上位 100 位までの類似度のランキングを作成し，検索に適合した小問文がどれだけ上位を占めているかを平均適合率により評価した．小問文が属する 24 の分野を元に 11 クラスを設定し，クエリが属するクラスと検索結果のクラスが同一であるならば，検索に適合しているとみなした．小問文が属する分野の情報については，東進ハイスクール[*5]の大学入試センター試験解答速報を参考にした．

小問文に含まれる自然言語文間のコサイン類似度 SIM_N を用いた手法 （以下 VSM）と，提案手法 （以下 VSM-MATH）による平均適合率の比較を行った．各クエリに関して上位 100 位までの平均適合率を計算し，そのクエリが属するクラス内における平均値を求めた結果を表 3 に示す．

クラス 1〜7 に関して，VSM-MATH が VSM を上回る結果となった．特にクラス“数列”に関しては，VSM では定義文の差異などが原因でクエリと適合しているにもかかわらず上位にならない小問が多かったが，$\{a_n\}$ など他のクラスで出現しづらい独特な数式構造が共通して用いられているため，数式間の類似度が考慮されたことで平均適合率が大幅に向上したものと考えられる．

表 3: 平均適合率の比較

クラス	分野	VSM	VSM-MATH
1	数と式，方程式，方程式と不等式，方程式と不等式・2 次関数，2 次関数	0.710	0.715
2	集合と論理	0.886	0.920
3	数列	0.639	0.842
4	指数，指数・対数関数，対数関数	0.584	0.622
5	微分法・積分法	0.698	0.749
6	ベクトル	0.852	0.964
7	コンピュータ，数値計算とコンピュータ，統計とコンピュータ，計算とコンピュータ	0.708	0.902
8	統計	1.000	1.000
9	平面幾何，図形と計量、平面図形	0.909	0.825
10	三角比，三角関数	0.702	0.560
11	場合の数と確率，確率，確率分布	0.812	0.738

∗3 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
∗4 http://www.w3.org/People/Raggett/EzMath
∗5 http://www.toshin.com

6. おわりに

本研究では，大学入試センター試験の数学問題文を対象とした検索を目標として，数式を含む小問文間の類似度を算出する手法を提案した．前提記述と質問記述の抽出実験では，少ない教師データから二つの記述を高い精度で識別できることを示した．小問文の検索に関する実験では，小問文に含まれる自然言語文のみを考慮した類似度を用いたときと比較して，クエリと類似した問題がランキングの上位により多く出現することが多くの出題で確認できた．

本研究では数式集合間の類似度に EMD を用いたが，二つの数式集合が包含関係にあった場合，それらが互いに等しいとみなしてしまうという問題点がある．数式の複雑さや集合の要素数などを用いて類似度に重み付けを行うなど，より実用に即した尺度を検討することが今後の課題である．

参考文献

[1] Angluin, D.: Finding Patterns Common to a Set of Strings, *Journal of Computer and System Sciences*, Vol. 21, pp. 46–62 (1980).

[2] Needleman, S. B. and Wunsch, C. D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, Vol. 48, pp. 443–453 (1970).

[3] 横井啓介, 相澤彰子: 類似性を考慮した数式検索手法の提案, 情報科学技術フォーラム講演論文集, Vol. 8, pp. 347–350 (2009).

[4] Miner, R. and Munavalli, R.: An Approach to Mathematical Search Through Query Formulation and Data Normalization, *In Towards Mechanized Mathematical Assistants*, Vol. MKM 2007, pp. 342–355 (2007).

[5] Kohlhase, M. and Şucan, I. A.: A search engine for mathematical formulae, *Proc. of Artificial Intelligence and Symbolic Computation*, Vol. number 4120 in LNAI, pp. 241–253 (2006).

[6] 市川宙, 橋本泰一, 徳永健伸, 田中穂積: テキスト構文構造類似度を用いた類似文検索手法, 情報処理学会研究報告, Vol. 2005-DBS-136, pp. 39–46 (2005).

[7] Rubner, Y., Tomasi, C. and Guibas, L. J.: The Earth Mover's Distance as a Metric for Image Retrieval, *International Journal of Computer Vision*, Vol. 40, pp. 99–121 (2000).