

行列表現を含む数式検索手法ならびに学習項目の抽出法の提案

進士晃太郎*1 宮崎佳典*2

静岡大学情報学部*1 静岡大学大学院情報学領域*2

1. 導入

数式には、分数や行列などといった、記号が水平に並ばない構造が存在する。こうした構造を含めた数式を計算機上で表現するデータ形式として MathML や TeX などが挙げられる。また、これらのデータ形式の普及に伴い数式のデータが蓄積されたことで、数式に対する検索技術の必要性が高まっている。これに対し我々研究グループは、MathML のタグセットの 1 つであり、数式の視覚的な構造を記述する MathML Presentation Markup に着目し、正規表現を用いた数式検索システムを開発してきた[1]。

しかし、現時点において、同システムは行列表現を含む数式に対応できていない。行列表現は主に線形代数学で用いられる、数学における重要な表現と言ってよい。線形代数学は多くの大学で微積分学と共に学ばれる、数学における主要教育科目でもある。これを踏まえ、本研究では同システムを拡張し、行列表現に対応できるように施すことを目的とする。さらにシステムの応用として、数式に含まれる学習項目を抽出する機能も実装する。この機能は学習支援システムとして用いることが想定され、学習者にとって初見の数式に対し、何を学習すれば数式を理解できるのか、ヒントを与えてくれる機能である。

数式検索の先行研究として、横井らの研究[2]や橋本らの研究[3]などが挙げられる。横井らの研究では、MathML Content Markup (数式の意味を記述するタグセット) に対して類似度を測ることで類似検索を行っており、橋本らの研究では検索対象の数式と Web 文書内に MathML Presentation Markup で記述されている数式の類似度を測り類似検索を行っている。

2. 既存システム[1]

本研究で扱うシステムは、MathML Presentation Markup を対象に検索を行うものである。その際、数式の (MathML) 表現のゆらぎを解消するために正規化処理を行うことで 1 つの数式に対応する MathML 表現を統一している。また、クエリ内に正規表現を許すことで曖昧検索機能を実現している。検索の結果、一致した部分をハイライト表示または置換を行うことが可能である。注釈までに、本検索システムでは数式の表示に関する情報を対象として検索を行

っており、数式の意味は考慮していない。

[1]の現在の入力インターフェースを図 1 に示す：

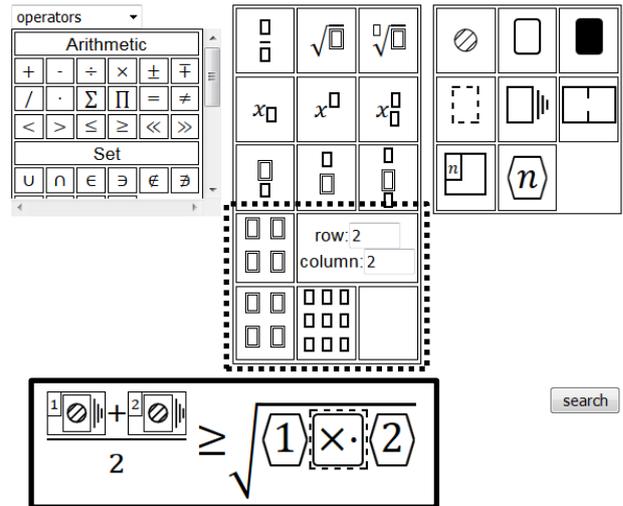


図 1: 数式検索システムの入力インターフェース

インターフェース左下部に、入力したクエリ用数式が表示される (図 1 の太線で囲まれた数式)。左上の記号群は数学で用いられる記号がまとめられており、クリックすればその記号が入力される。中央の 9 つのアイコンは分数や上下付き文字といった構造を入力するものである (破線で囲まれた部分は今回新規に追加したアイコンであるため 3 節にて後述する)。

右上の 8 つのアイコンは正規表現に係るものであり、以下にそれぞれ説明する通りである (表 1)。

表 1: 正規表現一覧

	任意の 1 文字		内部入力文字のいずれか		内部入力文字以外
	内部の文字列と最大 1 回マッチ		入力文字列の 1 回以上の繰り返し		仕切りごとの入力文字列のいずれか
	後方参照用のラベル (n)		n (番) を後方参照		

ユーザーは図 1 のパレット内コンポーネントを自由に組み合わせて複雑な曖昧検索を実現することが

Retrieving Mathematical Expressions Including Matrices and Extracting Mathematical Concepts

*1 Kotaro Shinshi, Faculty of Informatics, Shizuoka University

*2 Yoshinori Miyazaki, College of Informatics, Shizuoka University

可能となっている。クエリを作成し終えた後に、右下の search ボタンをクリックすることで左下に表示されている数式が検索されることとなる。図 1 の検索クエリには“相加・相乗平均”の式が入力されている(左辺の分子第 1 項(任意の文字列), 第 2 項(同)が右辺の平方根内で後方参照されている)。

3. 現行システムの行列表現への対応

現行数式検索システムを行列表現に対応させるためには、検索処理を行うコア部分に加え、それに付随する作業として、行列に対応した正規化を行い、入力インタフェースに行列を追加する必要がある。

MathML Presentation Markup における行列の表現には、行列状の構造全体を表す mtable 要素、行列の 1 行を表す mtr 要素、行列の 1 要素を表す mtd 要素が用いられる。mtr 要素は複数の mtd 要素を子に持つが、同 mtable 要素内の mtr 要素間で子を持つ mtd 要素の数が異なる場合、最も多くの子を持つ mtr 要素に合わせてその他 mtr 要素内の mtd 要素は左詰め表記される。すなわち、右側の任意個数の要素が空となる行列となる。このように空の要素を表現する場合、前述のように mtd 要素を記述しない場合と空の mtd 要素が記述される場合が考えられる。これは見かけ上は同じでも実際には異なる記述となるため、正規化の対象となる。よって、子の数が不足している mtr 要素に空の mtd 要素を追加することで正規化を行う。

mtd 要素は任意個の要素を子に持つ。mtd 要素の子として mrow 要素を記述し mtd 要素の子を 1 つにまとめる場合と mrow 要素を記述せずに 1 つ以上の子を記述する場合が考えられるため、こちらも正規化の対象となる。よって、mtd 要素の子を 1 つの mrow 要素に統一することで正規化を行う。

また、行列には二点リーダや三点リーダを用いて複数の要素を省略して表現する場合があるが、本研究ではこれらの省略表現はあくまでも二点リーダや三点リーダという記号として扱い、現段階においては、省略されている内容については考慮しないこととする。これは、二点リーダや三点リーダによって省略されている内容・範囲が書籍や執筆者に依存し、必ずしも再現できるわけでは無いためである。

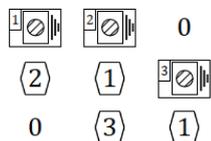


図 2: 本システムによる行列表現入力例

GUI (Graphical User Interface) についても行列を入力できるように新たなアイコンを追加する。新たな入力インタフェースは図 1 の破線で囲んだ部分である。中央のアイコン群の 4 行目のアイコンは、右側の row と column の欄に入力された値に従って任意の大きさの行列を指定するアイコンである。また、5 行目のアイコンは教科書などで頻出する 2×2 ならび

に 3×3 行列をそれぞれ表す。

今回の行列表現の対応により、作成が可能となったクエリの例を上図 2 に示す。これは対角成分がすべて同一な 3 次の三重対角対称行列を表す。

4. 数式に含まれる学習項目の抽出

学習項目を抽出する際は、数式検索システムで正規化された数式を対象とする。学習項目に対応する正規化された数式のパターンを定義しておき、それにマッチするかを現行システムで検索すればよいことになる。複数の学習項目によって新たな学習項目が生成される場合、各々の学習項目に合致すればよい。階層構造を構築してゆく(例:“行列方程式”は“行列”と“方程式”の学習項目に相当する数式パターンにマッチすればよい)。

学習項目の抽出機能は現在、行列に関連する学習項目の実装が進んでおり、行列、正方行列、対角行列、スカラー行列、単位行列、対称行列、上三角行列、下三角行列、三角行列、零行列の判定が可能となっている。これらは行列の 1 要素を表す mtd 要素を 2 次元配列に格納してチェックすればよい。また、行列以外では、対数、自然対数、常用対数などの学習項目判定が実装されている。図 3 に、入力した数式クエリと対応する学習項目を全出力した例を示す。

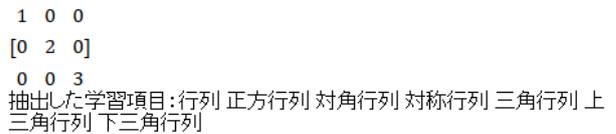


図 3: 学習項目の抽出機能の出力例

5. 今後の展望

本数式検索システムに関する今後の課題として、行列の検索機能の拡張や学習項目の抽出機能の拡張が挙げられる。行列の検索は可能となったものの、行列の大きさを不問とするような曖昧検索はできておらず、行列の曖昧検索を行う上では未だに自由度が低いと考えられる(行列の要素内については既に曖昧検索が可能である)。学習項目の抽出機能は現状では対応可能な学習項目の生成例が非常に少ないため、学習項目の拡張が最優先事項と言える。学習項目の出力方法についても改善の余地が認められる。

参考文献

[1]. 渡部 孝幸, 宮崎 佳典, 正規表現を用いた数式検索手法の提案, 情報処理学会論文誌, Vol.56, No.5, pp. 1417-1427 (2015).
 [2]. 横井 啓介 ほか, 数式構造と周辺テキストの両面を考慮した数式情報抽出, 情報処理学会, 第 73 回全国大会講演論文集, pp. 365-366 (2011).
 [3]. 橋本 英樹, 土方 嘉徳, 西田 正吾, MathML を対象とした数式検索エンジンの設計, 情報プロフェッショナルシンポジウム予稿集, pp. 45-59 (2007).