SURE 静岡大学学術リポジトリ
Shizuoka University REpository

正規表現を用いた数式検索

メタデータ	言語: ja 出版者: 静岡大学 公開日: 2016-06-15 キーワード (Ja): キーワード (En): 作成者: 渡部, 孝幸 メールアドレス: 所属:
URL	https://doi.org/10.14945/00009597

（課程博士・様式７）（Doctoral qualification by coursework,Form 7）

学 位 論 文 要 旨

Abstract of Doctoral Thesis

専　攻：情報科学専攻　　氏　名：渡部 孝幸

Course：Department of Information Science and Technology　Name：Takayuki Watabe

論文題目：

Title of　Thesis：

正規表現を用いた数式検索

論文要旨：

Abstract：

数式は, 数理的な概念やモデルを記述するために使われる. 数学はもちろん, 物理学や経済学など, 数理的な手法を用いるさまざまな分野で用いられる重要な形式である．web ページや電子書籍中に数式を記述したり, 数式の変形や数式のグラフの描画を自動的に行ったりなど, 計算機上での数式の利用も進んでいる．その結果，数式を検索するための技術の必要性が高まっている.

数式の検索に関する研究は複数行われている. それらの既存研究の多くは, 問い合わせとして入力した数式と検索対象となる数式との関連性の強さをもとに検索を行う．このような手法を用いると, 検索結果を順序付けて提示できる. この特徴は, 検索エンジンを実現する上で特に有用である. 一方, 問い合わせの数式の位置を特定することができない．このため,「数式中の，問い合わせが記述されている箇所を探す」といった処理を実現できない.

そこで本研究では, パターンマッチング（文字列探索）に基づく数式の検索手法を提案する. また, パターンマッチングの,「問い合わせ（パターン）の存在する位置を特定することができる」という特徴を利用し，検索した数式のハイライト表示や，数式に対する置換処理といった処理も実装する. 自然言語に対するハイライト表示は web ブラウザに搭載されている機能であり，数式に対するハイライト表示を行うことができれば，数式を含む文章を読む際に高い利便性が得られる．例えば，さまざまな関数について解説している web ページにおいて，正弦関数を分母に持つ関数について調べたい場合，その関数をハイライト表示できれば，高い利便性が得られるであろう．置換処理は，テキストエディタや文書編集ソフトウェアに搭載されている機能であり，数式の置換処理は，数式を含む文書を編集している際に効果を発揮する．例えば，2 人の著者が共同で文書を執筆しており，微分の表記を 1 人は$\frac{\mathrm{d}}{\mathrm{d}x}f(x)$，もう 1 人は$f'(x)$としていた場合を考える．ここで，$\frac{\mathrm{d}}{\mathrm{d}x}f(x)$という表記を$f'(x)$に置換できれば，表記を統一する作業の負担が軽減される.

本手法では，検索対象の数式の見た目が同一であれば同一の検索結果が得られるようマッチングを行う．一方，本研究で対象とする数式のデータ形式である MathML では，同一の数式を表示するためのデータが一意に定まらない．例えば，ab^2という数式を表すデータを作成する際，bの右上に上付き文字として 2 を配置することも，ab というまとまりの右上に 2 を配置することもできる．そこ

で，検索処理を行う前に MathML データに対して正規化処理を行うことにより，このようなデータのゆらぎを解消し，数式の表示が同一ならば検索結果も同一となるようなマッチングを実現した．

パターンマッチングの利便性を高めるために，正規表現を用いてパターンを記述する機能も実装した．正規表現における，ワイルドカード，記号の繰り返し，論理和や後方参照などの機能により，パターンの記述能力を飛躍的に高め，例えば「同一の分母を持つ分数同士の加算」といった数式をパターンマッチングで検索できるようにしている．ここで，数式のデータは木構造として表現されるが，正規表現には木構造を扱う能力がない．そこで本研究では，パターンを再帰的に処理できるよう拡張が施された正規表現エンジン Onigmo を利用することにより，正規表現で木構造を扱うことを可能とし，数式に対する正規表現を用いた検索を実現した．

正規表現を用いたパターンマッチングでは数式の意味を考慮した問い合わせを処理することができない．この問題を解決するための礎として，数式中の記号の意味推定を行う枠組みも提案および実装した．意味推定の例として，$|A|$という数式を考える．パターンマッチングにおいては，この数式は，‘|’，‘A’，‘|’という 3 つの記号が並んだものである．このため，「絶対値を検索する」，「行列式を検索する」，「集合の濃度（要素数）を検索する」といった問い合わせには対応できない．ここで，記号の意味を推定し，‘|’が「絶対値」「行列式」「濃度」のいずれの意味を持つかを判別できれば．意味的な問い合わせにも対応できるようになる．推定をする際の手がかりとして，「意味を推定する記号と同一の数式中に存在する他の記号」および「記号の順序」に着目した．例えば前者に関して，$|A \cap B|$という数式において，$\cap$は集合に対する演算を表すため，‘|’という記号が「集合の濃度」を意味すると考えられるであろう．このような推定の手がかりを数値化した上で，機械学習を用いることにより，意味の推定を行った．