MathML で記述された数式に対するうろ覚え検索手法

長尾　悠真

Web 上の情報は莫大であるため，必要な情報を探し出すためには検索が必須である．また，Web ページ上の情報は文章ばかりではない．例えば，数学や物理などについてのページには数式が記載されている．Web ページに数式を記述する際にはプレーンテキストを用いることがあるが，プレーンテキストで記述された数式を検索することは困難である．これは，数式は独自の構造を持っており，従来の検索エンジンなどによるテキストのマッチング方法ではその構造を捉えることが難しいことが原因である．

一方，2014 年の 10 月に W3C 勧告された HTML5 で MathML (Mathematical Markup Language)が部分的に導入された．これにより，MathML を用いた数式を含む Web ページが今まで以上に増加すると期待されている．特に，MathML の Presentation Markup 形式は数式の構造を表現することができ，プレーンテキストの数式よりも表記ゆれが小さいため，プレーンテキストのものよりも適切に検索が行えると考えられる．

うろ覚えの数式があった場合，公式名などを覚えていれば，検索エンジンなどで公式名をクエリとすることで目的の数式を含む Web ページを発見することができる．しかし，公式名などを覚えていない，または式に名前がついていない場合など，クエリを表現できず検索が困難な場合も少なくない．このような場合，数式をクエリとした検索を行うことができれば有用であると考えられる．そこで本研究では，Web ページ上の Presentation Markup 形式で記述された数式に対するうろ覚え数式検索手法を提案する．

提案手法では，クエリの入力は既存の数式エディタを用いて行えるようにしている．入力の際，正確に覚えていない部分にはワイルドカードを指定することができる．その後，クエリを MathML の Presentation Markup 形式に変換し，検索対象の MathML 式との比較を行う．なお，MathML 式は表記ゆれを含むので，比較に先立って表記ゆれの解消処理を施している．また，数式の比較にあたっては，数式の覚え間違いなどにも対応するため，クエリに完全に一致するものだけでなく，ノード列に対する編集距離を用いた数式間の類似度計算を行っている．検索結果は編集距離に基づいてランキングしたものを表示する．

提案手法の有用性を評価するため，提案手法を Ruby を用いて実装し，表記ゆれの解消や処理時間等に関する評価実験を行った．その結果，提案手法によって MathML で記述された数式に対するうろ覚え検索が概ね適切に行えることが示された．

（指導教員　鈴木伸崇）