木編集距離を用いたMathML数式検索手法

根本　孝徳

数式は数学・理工学など多くの分野で公式や定理などを記述するために用いられ，Web上で数式を使用するケースも多い．しかし従来的なテキストベースの検索エンジンでは，数式をクエリとして適切な検索を行うことは困難である．数式の検索においては単にテキストを比較するだけではなく，数式の意味や構造も考慮する必要がある．

一方，Web 上での数式利用を目的としたマークアップ言語である MathML(Mathematical Markup Language)が普及し始めている．MathML は数式の持つ意味や構造を記述することができるため，これを用いることで数式をクエリとして文書を適切に検索することが可能であると考えられる．一般に，クエリと完全一致する数式のみを検索する場合には，単純にソースコードを比較すればよい．しかしクエリと同じ意味を持つ式であっても，レイアウトや装飾のような数式の意味とは無関係の検索ノイズを含む場合や，数式の概形しか分からないためクエリが曖昧である場合など，単純な比較での検索は不可能であることが多い．したがって，何らかの手法でクエリと意味や構造的に類似する数式の検索を実現する必要がある．本研究では，MathML で記述された数式を対象とし，数式の意味や構造を考慮した数式検索手法を提案する．

木構造は XML を始め様々なデータを表現する際に用いられ，木構造間の類似度を求める方法として木編集距離が知られている．2 つの木間の木編集距離は，2 つの木を一致させる際に必要な編集操作(ノードの削除，挿入，置換)のコストの合計によって表される．本手法は MathML 数式のソースを構文解析して木構造を取得し，クエリとの木編集距離を求めることによって数式の類似度を判定する．木編集距離計算に用いる編集操作のコストを，数式を構成する各要素の意味や構造を考慮して定義することにより，クエリと意味や構造的に近い数式の検索を行う．また，木構造を解析して意味的に重要ではない要素を取り除き，検索ノイズの除去を行うことによって，より精度の高い検索を実現する．

以上の提案手法を Ruby で実装し，実際に Web ページ上で使用されている MathML 数式を用いて評価実験を行った．その結果，本手法を用いることでクエリと完全一致する数式だけなく，クエリと類似する数式を検索可能であることが確認された．また，木編集距離と数式の類似度にある程度の相関関係がみられ，検索結果の適合率という点では木編集距離を数式の類似度として用いることの有効性が示された．今後の課題としては，編集コスト定義の見直しによる更なる精度向上や検索処理の高速化などが挙げられる．

（指導教員　鈴木伸崇）