DTD を用いた XML キーワード検索システム

進藤　千愛

XML(Extensible Markup Language)は，さまざまなデータを管理・活用するためのフォーマットとして広く用いられている．XML 文書内を検索する手法も多く提案されている．これらの手法は，XML キーワード検索と XQuery などの問い合わせ式で検索する手法に大別される．前者はユーザが任意の検索語(キーワード)を用いて XML 文書を直接検索する．検索結果として，キーワードをすべて含む部分文書が返される．条件を満たす(キーワードをすべて含む)部分文書を得るために，LCA(Lowest Common Ancestor)という特定のノードを XML 文書から抽出する手法が提案されている．LCA とは，キーワードが含まれるノードの共通祖先の集合のうち，最も子孫に位置するものである．この手法の場合，検索結果は LCA に該当するノードを根とする部分文書である．

XML キーワード検索は，検索に際して特別な専門知識が不要である点が長所である．しかし，検索に用いたキーワードには曖昧性が存在する．すなわち，検索対象が要素ノードかテキストノードかの区別をしないため，ユーザが意図した検索結果を得られない可能性がある．これに対して，問い合わせ式による検索では，XML 文書の構造に基づいて詳細な条件を記述して検索する．したがって，意図した結果を得やすいという長所がある．その反面，検索のために(1)構造に基づく要素ノードの指定，(2)検索条件に該当する関数の選択・記述を要する．専門知識，文書構造の把握，適切な関数の選択に加え，検索条件が複雑(詳細)になれば記述に要する手間も相当かかることが考えられる．

そこで，本研究ではユーザから与えられる条件と LCA 用いて XQuery 式を生成する手法を提案する．条件は，キーワードの曖昧性を回避するために，簡単な式(キーワード式)の形で入力する．また，LCA は先行研究とは異なり，DTD から抽出する．したがって，ユーザは DTD で使用されている要素がわかれば，文書の構造を完全に把握していなくても XQuery 式を生成できる．キーワード検索に関する先行研究はいくつか存在するが，XQuery 式生成に DTD を利用しユーザ側の負担軽減を図ったものは著者の知る限り存在しない．

上述の提案手法を Ruby で実装し，評価実験を行った．評価実験では，様々な組合せでキーワード式を構成し，キーワード式と生成された XQuery 式と記述量の比較を行った．いずれの組合せにおいてもキーワード式の方が記述量が大幅に少ないという結果が得られた．生成 XQuery 式は構造上不要な部分は除去された状態であったことから．提案手法の有用性が確認できたと考えられる．

（指導教員　鈴木伸崇）