論 文 内 容 の 要 旨

博士論文題目

A Study on Effective and Efficient XML Element Retrieval Considering Document Updates

（文書の更新を考慮した高精度かつ高速な XML 部分文書検索に関する研究）

情報科学研究科情報科学専攻　欅 惇志

（論文内容の要旨）

W3C によって策定された Extensible Markup Language (XML) はデータ交換の標準フォーマットとして広く利用されており，近年は様々なアプリケーションのデータ，Wikipedia 記事や Office 文書など多くの用途に利用されている．そのため，現在までに膨大な数の XML 文書が蓄積され，今後ますます多くの XML 文書が作成されると予想される．

従来の検索エンジンのうちのほとんどは，検索結果として文書のリストを提示する．そのため，ユーザは各文書中から欲しい情報を自ら発見する必要があるが，記述量の多い文書から求める情報を抽出する際には大きな労力を必要とし，また，長時間を費やしても欲しい情報を含まない可能性もあり，これらの作業はユーザにとって大きな負担である． それに対して，XML 部分文書検索では，検索結果として文書のうちユーザが必要とする情報が記述された箇所を特定して提示することを目指す．従って，ユーザの検索時の労力を軽減することが可能な，非常に有用な検索技術である．

既存の XML 部分文書検索に関する研究では，主に，1) ユーザが求める必要十分の内容を検索結果として提示することを目指す高精度な検索と，2) ユーザに対して高速に検索結果を提示することを目指す高速な検索が取り組まれてきた．我々は高精度な検索を実現するために，情報量の多い部分を発見するためのスコアリング手法と，文書中の検索結果として最適な部分を特定するための検索結果構築手法の提案を行った．評価実験の結果，提案手法は従来の手法と比較してより高精度に検索が可能であるという結果が得られた．

また，我々は検索システムの実運用を想定した場合には必ず発生する文書の更新への対応を目指した．なぜなら，これら文書の更新に対応しなかった場合には，検索システムはユーザに対して適切な検索結果を提示することはできず，その結果，検索システムの利便性が低下するためである．我々は，一般的な構成の XML 部分文書検索システムに差分更新機能を搭載させるべく新たな索引構造を定義し，更新対象のうち不要なデータを除外するためのフィルタを提案した．更に，索引語の重み算出時において正確な大域的重み算出のためのパス式統合手法の提案を行った．評価実験の結果，提案手法を適用することで，検索精度の低下を抑制しつつ，極めて短時間で索引の更新を実現した．

これらの研究の成果物として，正確な検索と高速なクエリ処理，そして文書の更新の即座の反映を満たした実用的な XML 部分文書検索システムを開発した．XML 部分文書検索技術の期待される応用対象として Web 文書が存在する．XML 部分文書検索と，Web 文書の中でも代表的なファイルフォーマットである HTML 文書に対する部分文書検索の間に存在する差異を解消するため，HTML 文書に対して部分文書検索技術の適用を目的とした文書の整形手法の提案を行った．

氏　名	欅 惇志

（論文審査結果の要旨）

平成25年12月19日に開催した公聴会の審査結果を踏まえ，平成26年2月21日に本博士論文の最終審査を行った．その結果，本博士論文は，提案者が独立した研究者として研究活動を続けていくための十分な素養を備えていることを示すものと認める．

欅惇志は，本博士論文において，文書の更新を考慮した高精度・高速な XML 部分文書検索手法を提案し，その有効性を示している．本論文の具体的な貢献を以下に示す．

高精度な XML 部分文書検索の実現に関して，従来の研究においては文書検索の検索手法を基にして部分文書検索への拡張を行ってきた．しかし部分文書検索では単に検索質問に対する適合箇所を含む文書を発見するだけではなく，適合箇所そのものを特定する必要があるため，単に文書検索手法を拡張するだけでは不十分である．そこで，1）重要部分抽出の要件を考慮したスコアリング手法，2）文書中の最適粒度発見のための部分文書統合手法，3）関連する部分文書の統計量を考慮したスコアリング手法を提案し，本手法の有効性を明らかにしている．

Web 上の文書をはじめ多くの電子化文書は更新が頻繁に発生することは通常であり，これらの更新に対応しなければ，検索精度の観点からその検索システムの有用性は次第に低下する．これに対して，従来の XML 部分文書検索システムでは文書の更新を想定していないため，更新に対応するためは長時間を要した．そこで，本論文では 1）高速な索引語の重み算出を実現するための索引構造，2）不要なデータを特定するためのフィルタ，3）少数のデータから高精度な統計量を算出する手法を提案し，高速な更新処理と高い検索精度を両立可能であることを実証している．

上記の研究成果から，高性能かつ高精度な XML 部分文書検索システムを実現可能としたことに付け加え，本技術をより一般的な構造化文書である Web 文書，すなわち HTML 文書への適用を試みている．そのために，1）文書内容の論理構造と文書の物理構造一致のための再構造化，2）不要な箇所を特定するためのフィルタを提案し，情報の欠片を検索する質問に対して有効であることを示している．

このように，実用的な XML 部分文書検索技術ならびにその一般的な Web 文書への展開に関する研究において，新たなすぐれた技術を提案した本論文は，工業的な価値だけでなく情報検索分野に対して学術的に大きく貢献したものと評価できる．よって，本論文は，博士（工学）の学位論文として十分な価値があるものと認める．