

氏 名	し 清 水 とし 敏 之
学位(専攻分野)	博 士 (情 報 学)
学 位 記 番 号	情 博 第 297 号
学位授与の日付	平 成 20 年 3 月 24 日
学位授与の要件	学 位 規 則 第 4 条 第 1 項 該 当
研 究 科 ・ 専 攻	情 報 学 研 究 科 社 会 情 報 学 専 攻
学 位 論 文 題 目	A Study on Document-Centric XML Search (文書指向XMLに対する検索に関する研究)

論文調査委員 (主 査) 教 授 吉 川 正 俊 教 授 石 田 亨 教 授 田 中 克 己

論 文 内 容 の 要 旨

現在、XMLは電子化された情報の標準的な記述フォーマットのひとつとして広い範囲で使われている。XMLの普及に伴い、XMLの変換、連結などの関連技術も開発され、もはや、XMLはそれを利用している意識がない場合でも、システム内部では利用されているといっても過言ではない。XMLはタグで構造化されたテキスト文書であり、その論理的な構造は木である。論理構造を持つ文書は一般的にXMLでも表現することが可能であると考えられるため、標準的に利用可能であるXMLに関して考えることにより、構造化文書一般に関しても技術が応用可能であると思われる。

本論文では学術論文や本などをXML化した、いわゆる文書指向XMLに対する検索に着目している。文書指向XMLは自然言語文による長いテキストを含んでおり、スキーマが複雑である傾向があるため、キーワードを用いた検索の要求が強いと考えられる。問合せにキーワードを含め全文検索を行うことや、キーワード集合を基にした問合せにより情報検索を行うことが考えられる。XML全文検索では構造条件と含まれるべきキーワードを組み合わせて指定し、適合する要素を取得する。XML情報検索ではキーワードを基にした問合せに対して、その検索要求に適合するXML文書中の要素を適合度付き、もしくは順序づけして取得する。本論文ではXML文書に対する全文検索や情報検索に関して研究を行っており、研究の成果は以下の三項目にまとめることができる。

第一に、XML文書に対する全文検索のためのB+木索引の提案である。XML文書の全文検索に関しては、W3Cにより全文検索に関するワーキングドラフトが発表され、注目され始めているが、その高速化に関しては、まだ十分な研究が行われていない。本論文ではテキスト片と逆経路情報を含む識別子の組合せをエントリとするB+木を構築することにより、XML文書に対する全文検索の高速化を実現している。出現一意性を考慮したテキスト片の切り出しを行うことで必要十分なテキスト片のみを保持し、逆経路情報を利用して探索キーを順序づけすることで同じタグ名の要素の効率的な取得を可能にしている。そして、実験において先行研究と比較した問合せ処理の大幅な高速化を確認している。

第二に、関係データベースを利用したXML情報検索システムの開発である。XML情報検索システムでは、XML文書の構造を利用することで部分文書を対象とする検索が可能となる。問合せにキーワードを用いることで、XPathやXQueryといったXML専用の問合せ言語を用いることなく、従来のWEB検索エンジンのような直観的な検索を利用者に提供する。提案するKikoriシステムは、関係データベースを用いて実装されており、実用に耐える検索速度を実現している。また、XML向けの索引語の重み付け式や検索結果提示インタフェースに関しても考察を行っている。利用者インタフェースとして、結果要素を文書単位にまとめ、その中で要素を文書順に提示するFetchHighlightインタフェースを提案しており、さらに検索結果要素を実際の文書のページレイアウト上に重畳して提示することにより直観的な検索結果の把握を助けることを考えている。

第三に、XML情報検索における利得と閲覧コストに基づく検索結果の取得と評価の提案である。提案するシステムでは、利用者は支払ってもよい閲覧コストの合計量を指定し、それに対してシステムは柔軟に入れ子のない要素集合を取得する。

XML情報検索では取得する要素の大きさは多様性に富み、検索結果は入れ子する可能性があるが、閲覧コストの概念を導入することで検索結果の量を制御し、同じ閲覧コストの中でより多くの利得を取得できるように要素を取得することで入れ子を除去した要素集合を結果として取得する。検索結果の取得アルゴリズムとして、入れ子を考慮しつつ貪欲に利得獲得効率のよい要素を取得していくものを提案し、さらに、取得可能な利得の上界値と比較することによるシステムの評価手法を提案している。

論文審査の結果の要旨

本論文はXMLに対する検索、特に文書指向XMLに対する検索として近年着目されているXML全文検索やXML情報検索を扱っている。XMLは広く利用されており、構造化文書一般にも提案技術が適用可能であることを考えると、本論文の技術の適用範囲は広い。

XML文書に対する全文検索のためのB+木索引の提案では、XML木のノードに文書構造を表現するラベルを付与し、木構造の経路に対応するラベル列と対応する文書片を探索キーとして採用している。高速化と容量の圧縮を実現するために、探索キーの順序づけや探索キーとして保持する情報に工夫を行っている。B+木は多くのデータベースで利用されているため、それを利用した索引は広く普及する可能性がある。XML全文検索を高速化する索引に関する研究は少なく本研究は新規性が高い。また、実際に索引を実装し先行研究と比べた問合せ処理時間の優位性を示しており、有用性の点でも評価できる。

関係データベースを利用したXML情報検索システムの開発では、末端利用者の視点に立った利用者インタフェースの提案とその開発を行っている。実際に動作し、XMLの専門家でない利用者でも利用可能なシステムを汎用性の高い関係データベースを利用して構築したことは評価できる。また、XML情報検索の国際的なプロジェクトであるINEXに参加し、システムの検索結果の適合性に関し良好な結果を残しており、システムの有効性を示した点も評価できる。

XML情報検索における利得と閲覧コストに基づく検索結果の取得と評価の提案では、利用者の閲覧コストの量に対して柔軟に検索結果を取得してくる検索方式を提案している。このような検索方式はこれまでにない新規性の高いものである上に実用性も高いものである。また問題の定式化を行い、与えられた閲覧コストに対する利得の上界を理論的に求めることによりシステムの検索性能の評価に利用している点も評価できる。

このように、本論文は、文書指向XMLに対する検索に関し、新たな索引構造の提案と実装、実用性の高いシステムの構築、新たな検索方式の提案と解析を行ったものであり、研究の有用性と実用性が高く評価された。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。

また、平成20年2月6日実施した論文内容とそれに関連した試問の結果合格と認めた。