平成27年度 情報工学コース卒業研究報告要旨

関 研究室	氏　　名	西 村　卓
卒業研究題目	TopTreeによる木圧縮法の実装および直接問合せ処理法の提案	

近年，クラウドサービスの普及やwebコンテンツの発展などにより，大量のデータの蓄積や伝達が必要不可欠となっている．データ交換のための様々な文書形式が提案・運用されているが，その中にXMLが挙げられる．XMLは最も広く用いられているマークアップ言語であり，使用者はタグを用いて文書やデータの構造を比較的自由に表現することができる．一方で，実用的なXML文書においては，そのタグ構造を反映した木構造として扱うと，データサイズがしばしば莫大となる．これらの大規模なXML文書は圧縮して保存する必要がある．

XML文書の圧縮法として，文書を文字列とみなして圧縮する方法と，文書の構造を木表現としたものを圧縮する方法とがある．木表現の圧縮法は，圧縮データに構造情報が一部保持されるため直接操作が可能であるという利点があり，本研究でも，木の圧縮法に着目する．木を圧縮する様々な方法がすでに提案されており，最も単純な方法として共通部分木を1つにまとめることで圧縮するDAG表現による方法が挙げられる．しかしこの方法では，部分木となっていない共通の内部構造(木の中間部分)をまとめることができない．これが可能である方法の1つとしてtop treeと呼ばれるデータ構造を用いる圧縮法が提案されている．ある木tに対するtop tree τとは，各頂点がもとの木tのクラスタと呼ばれる内部構造を表し，各辺がtのクラスタ間の併合関係を表すような2分木である．木tの最小のクラスタは辺であり，これはtop tree τの葉頂点に対応する．top treeによる木の圧縮法では，圧縮対象の木をtop treeに変換した後，さらにそれをDAG表現に変換する．この手法は圧縮率の自明でない下界が知られている等の特長をもつが，実際の圧縮性能に関する研究はほとんどなされていなかった．

本研究では，まず，XML文書の木表現をtop treeに基づき圧縮するツールを実装した．特に，木に対するtop treeは併合するクラスタ対の選択順序によって構造が変化し一意に定まらないため，選択順序を決定する2つのアルゴリズムを提案し実装した．具体的に，簡単な基準でクラスタ対を併合する素朴なアルゴリズムと，出現数の多い隣接クラスタ対から優先して併合するRePairアルゴリズムを実装した．さらに，実用的なXML文書は圧縮及び解凍に大きなコストがかかるため，top treeに基づく圧縮データをいったん解凍することなく直接問合せ処理する手法を提案し実装した．具体的に，問合せを表す決定性選択トップダウン木オートマトンと圧縮XML文書を入力とし問合せ結果である選択頂点の集合を得るツールを実装した．

作成したツールを用いて様々なXML文書に対して圧縮と直接問合せを行った結果，多くの場合RePairアルゴリズムの方が素朴なアルゴリズムよりも圧縮率は向上するが，圧縮に要する時間は大きくなることが分かった．また，圧縮アルゴリズムの種類は直接問合せの時間に対する影響は少ないことにが分かった．