平成16年度　コース卒業研究報告要旨

吉川 研究室	氏　　名	井　ノ　口　伸　人
卒業研究題目	文書スキーマを利用したXMLファイル編成法に関する研究	

XML 文書は，データ交換のためのフォーマットとして様々な分野で用いられるようになっている．その一因として，文書の取り得る構造に関する文法を記述できる点が挙げられる．その文法を記述するための言語が，DTD, W3C XML Schema, RELAX NG などのスキーマ言語である．DTD は，XML の仕様書 Extensible Markup Language (XML) 1.0 中で規定されているスキーマ言語であるが，名前空間をサポートしておらず，データ型に関する情報も記述できないという二つの大きな問題が存在する．W3C XML Schema と RELAX NG はそれぞれ W3C と OASIS が制定したスキーマ言語であり，これらはどちらも DTD が持つ二つの問題を解決している．データ型に関する情報は，データの物理的配置を決めるために重要なものである．そこで，本論文では，ネイティヴ XML データベースに文書を格納する際に，スキーマ文書から得られる情報を利用することで，一般性を失わずに，データベースごとに異なる最適なファイル編成を行う方法について考える．既存の研究では，索引付けの対象は主に単一の XML 文書であり，また，複数の XML 文書を対象としている場合でも各文書に ID を振るだけであるなど，単純なものが多い．しかし，単一のスキーマ文書について多数の妥当な XML 文書が存在し，それらをネイティヴ XML データベースに格納することを考えたとき，XML 文書の間の関係は無視することはできない問題になる．そこで本論文では，XML 文書の間の関係について考察し，格納時に考慮すべき点について述べる．

効率的な格納のために，スキーマ文書が表現する構造から，グループ化可能なものを抽出する．本論文では，グループ化可能な経路パターンとして，Temopral Branch と再帰循環を挙げた．これらによってグループ化された要素群は，根からの経路が完全に一致してはいないが類似しており，表現しているものを同列に扱うことが出来る．次に，XML 文書中のオブジェクトと，コンテンツを定義する．オブジェクトは ER モデルの実体を表現し，コンテンツは実体の属性を表現しており，それぞれ関係のレコードと属性に対応する．そして，複数のオブジェクトを含む XML 文書の中から，オブジェクトとコンテンツを特定する手法について提案する．オブジェクトまたはコンテンツを表現するノードを，経路パターンによってグループ化し，下図のように同一パターンのノードを連続領域に配置することで，目的の部分文書へのアクセスを効率化する．