4N-5

確率的スキーママッチングによる複数XMLデータに対する問合せ

添野 隆之† 天笠 俊之‡ 北川 博之‡

†筑波大学情報学群情報科学類 ‡筑波大学大学院システム情報工学研究科

1 はじめに

XML (Extensible Markup Language)[1] とは, データの意味や構造を記述するためのマークアップ言語の一つである. XML を利用することで様々なデータを柔軟に表現することが可能であるため, 近年, 多くの分野で利用され普及している. また, XML の普及に伴い, 複数の異なる XML スキーマに対するスキーママッチングを用いたデータ統合や問合せ処理の効率化の重要性が高まっている。

XML におけるスキーママッチングは複数の異なるスキーマに基づく XML データの統合を目的とした手法である. まず, ターゲットスキーマ T, ソーススキーマ S をマッチングし, T と S の要素間の対応関係を発見する. 発見された対応関係を基に複数の可能マッピングを生成する. 生成された可能マッピングを利用することで, T に基づくデータを S に基づくデータに変換する, 問合せを書き換えるといったことが可能となる. また, 可能マッピングはそれぞれ確率が与えられている. このため, 確率を考慮することでより信頼性の高い結果を得ることが可能である.

XML スキーマに対するスキーママッチングには次の問題点がある.

- 非常に多くの可能マッピングが生成されるため, 不確実性が大きくなる.
- 考慮すべき可能マッピングの数が増加し, 問合せ処理の効率が悪化する.

これらの問題に対して, Reynold ら [2] は膨大な数の可能マッピングを扱うための新たなデータ構造と, そのデータ構造を利用し確率を考慮しながら問合せ処理を行う手法を提案している. しかしながら, 一つのターゲットスキーマに対し単一のソーススキーマを想定しており, 複数のソースが存在する場合, 1 対 1 のマッチングを複数繰り返すことになり非効率である.

そこで本稿では, Reynold らにより提案された手法を拡張し, 複数のソーススキーマが与えられた場合のスキーママッチングを利用した問合せ処理を行う手法について提案する.

Query Processing against Multiple XML Data using Probabilistic Schema Matching
Takayuki SOENO†(soen@kde.cs.tsukuba.ac.jp)
Toshiyuki AMAGASA†(amagasa@cs.tsukuba.ac.jp)
Hiroyuki KITAGAWA‡(kitagawa@cs.tsukuba.ac.jp
†College of Information Sciences, University of Tsukuba
‡Graduate School of Systems and Information Engineering, University of Tsukuba

2 関連研究

ここでは, 本研究において重要な役割を担うブロック木について述べる.

ブロック木とは, Reynold らにより提案された, 可能マッピングを効率よく管理し, 利用するためのデータ構造である. ブロック木の構築は次の手順で行われる.

1. 各可能マッピングで共通する部分をもつものを, ブロックにまとめる.
2. 生成されたブロックについて, 一定のしきい値以上のブロックを, 制限ブロック (c-ブロック) とする.
3. c-ブロックを利用して, ブロック木を構築する.

構築されたブロック木は, ターゲットスキーマと同じ構造を持ち, 各ノードには 0 個以上の c-ブロック連結リストが接続されている. また, ブロック木には, c-ブロックのみが保持されているため, 省スペースで可能マッピングを管理することが可能となる.

3 提案手法

本研究で提案する手法について, 図 1 から 3 に示されるスキーマ T, S_1, S_2 を用いて説明する. まず, T と S_1, T と S_2 それぞれでスキーママッチングを行うと, 図 4, 5 に示される可能マッピングがそれぞれ生成される. 次に, 可能マッピングを利用しブロック木を生成する. 生成されたブロック木の一部を図 6 に示す. ここで, 図 4 の m_{11}, 図 5 の m_{21} に着目すると, どちらも book〜book を持っている事がわかる. これを利用しブロック木を変形すると, 図 7 に示すブロック木が得られる.

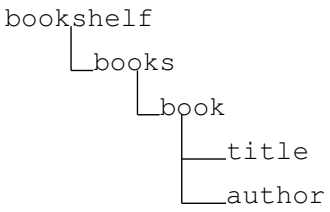

図 1: ターゲットスキーマ T

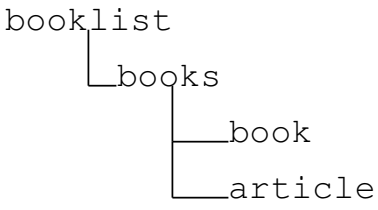

図 2: ソーススキーマ S_1

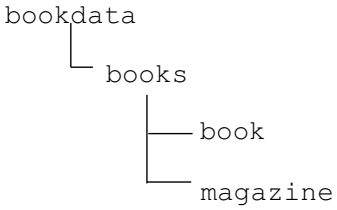

図 3: ソーススキーマ S_2

このように, 共有可能なブロックを 1 つにまとめて保持することで, さらに効率よく可能マッピングを管理することが可能となる.

次にこのような問合せを考える.

$$Q = \text{/bookshelf/books/book}$$

Q の書き換えを行う場合, 図 6 のブロック木よりも図 7 のブロック木の方が考慮すべきブロックの数が少ないため, より効率よく処理を行うことが可能となる.

4 まとめと今後の課題

Reynold らにより提案されたブロック木を拡張し, 複数の XML スキーマが与えられた場合の問合せ手法を提案した. 今後は提案手法の実装と, 人工データ, 実データを使った評価実験を行う予定である. また, ブロック木構築の最適化手法の検討を行う予定である.

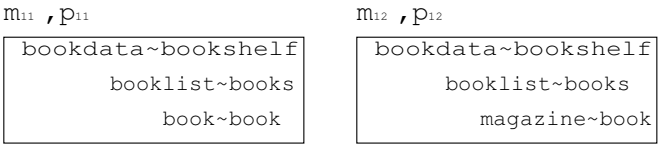

図 4: T と S_1 のマッピング

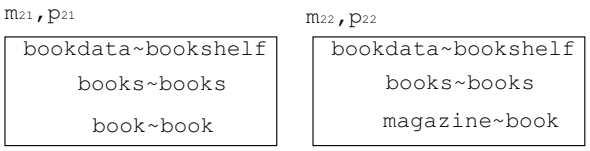

図 5: T と S_2 のマッピング

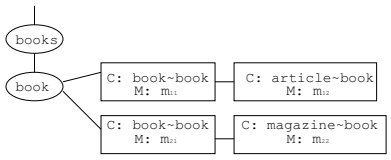

図 6: ブロック木

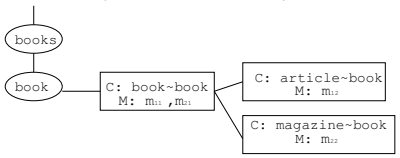

図 7: 提案手法により生成されたブロック木

謝辞

本研究の一部は科学研究費補助金 (#21650017) による。

参考文献

[1] http://www.w3.org/TR/REC-xml/

[2] Reynold Cheng, Jian Gong, David W. Cheung. Managing Uncertainty of XML Schema Matching. ICDE 2010, pp. 297-308, Long Beach, California, USA, March 1-6, 2010