ロバスト構造化文書処理技術

武市正人　胡振江　筧一彦

情報理工学系研究科数理情報学専攻

概 要

ロバスト構造化文書処理技術プロジェクトは，Programmable Structured Document（PSD）の手法を提案し，堅牢な文書処理技術の方法論を構築することを目標としている．PSD プロジェクトは，XML に代表される構造化文書のなかにプログラムの記述を許し，このプログラム記述に対して，これまでに得られている関数プログラミングおよびプログラム変換手法を適用することにより，ロバストでかつ効率的な構造化文書処理を実現しようとするものである．以下では，本年度の成果を，アプリケーション事例，PSD 基盤技術の開発，双方向計算機構の三点から説明する．

1　はじめに

データや文書を表現する方法として，XML（eXtensible Markup Language）と呼ばれる，構造を表すためのタグが付けられたテキストが広く普及するようになった．例えば学校の成績表であれば，学年の下にクラスが，クラスの下に学生が列挙され，その学生の下に科目毎の成績が示される，というような構造を持つことになる．このような電子的な構造化文書情報の蓄積と効果的な情報利用技術は，インターネットを含む広範な情報の交換・流通にとってきわめて重要な位置を占めている．XML に代表されるこれらの技術は，発展の著しい WEB による情報環境に向けて既存の技術の延長線上で実務的に開発されたもので，事実上の標準となってはいるがその言語的な概念が十分には整理されていない．このような体系的な処理技術の欠如が今後の情報交換の発展を阻害し，

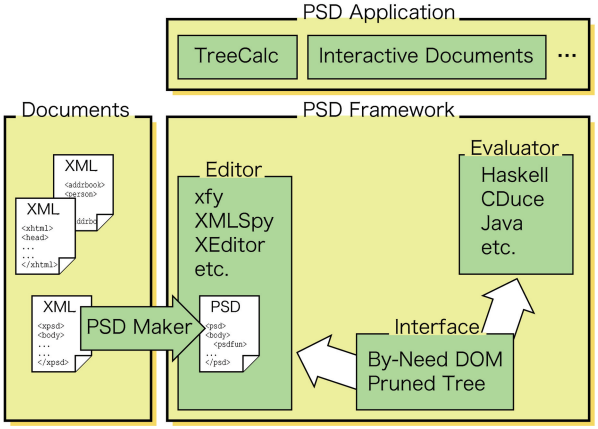

図 1: PSD プロジェクト概要

既存技術による個別的対処や人手による個別対応では一般性を欠く文書情報を蓄積することとなっており，この問題を解決することが重要な課題となっている．

構造化文書はプログラミング言語のデータ構造と類似しており，文書処理のアルゴリズムを記述するためには関数型言語が適している．Programmable Structured Document（PSD）は，プログラムの記述を含む文書を対象とし，これまで関数型言語での処理および変換手法を利用して，ロバストでかつ効率的な構造化文書処理を実現しようとするものである．すなわち，構造化文書をプログラミングにおける構造化データであるとみなし，プログラミング言語に関する理論を適用することによって，安全かつ信頼性の高い処理を実現する．また，処理を行うコードを対象文書に埋め込むことで，文書の高い可搬性を実現する（図 1 参照）．

PSD 実現のために必要となる基盤技術は，

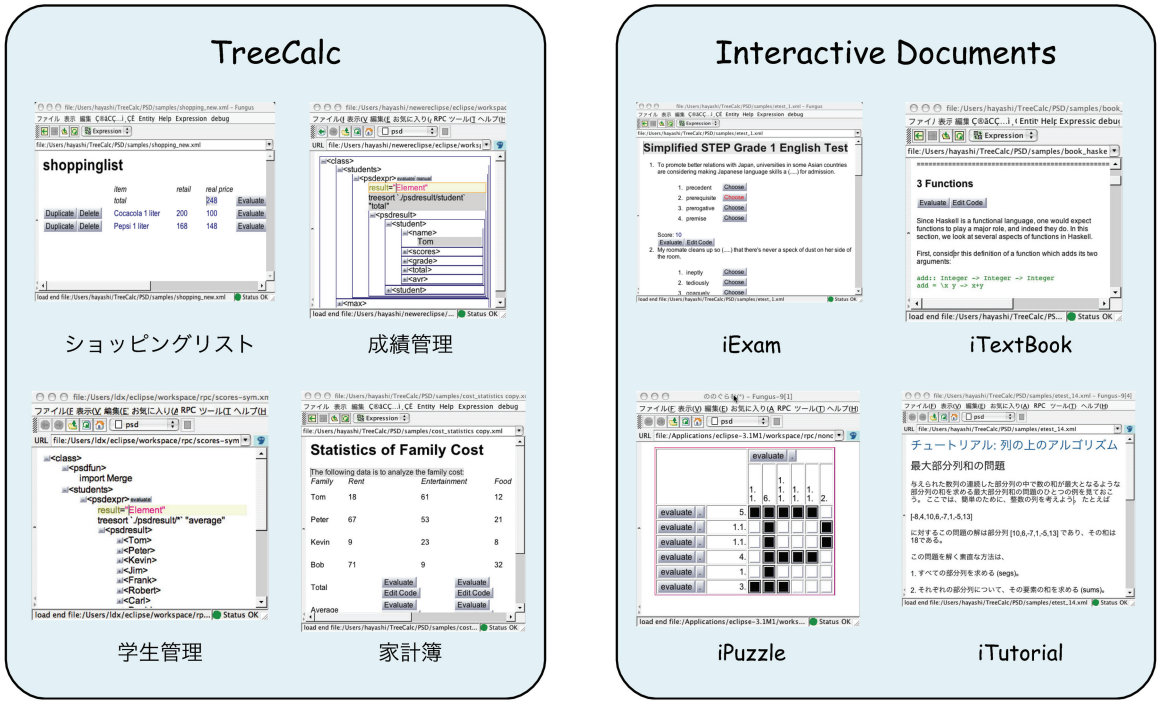

図 2: アプリケーション事例

- 構造化文書に必要な準構造データの概念を型として捉える形式的枠組の定義および，データ型に基づく効率的変換手法，
- PSD のための計算機構を組み込んだ構造化文書の実現手法，
- 関数型言語におけるデータ型の理論を発展させ，準構造データに適した型の理論を構築（これは，代数的プログラム変換，すなわち運算の成果を構造化文書に適用し，自己参照による文字変換や変換戦略を文書自体に付随させるという運算随伴機構に関する理論を含む）

の三つである．

2 平成 16 年度の成果

本年度の成果を，アプリケーション事例，PSD 基盤技術の開発，双方向計算機構の三点から説明する．

2.1 アプリケーション事例

構造化文書中にプログラムコードを保有する PSD の有効性を実証するため，さまざまなアプリケーションの実装を行なっている．これらは図 2 が示すように，主に表計算を木構造へと拡張した TreeCalc，それからユーザの入力に応じて構造化文書に埋め込まれたコードが結果を返すことによって実現される Interactive Document である．

2.2 PSD 基盤技術の開発

前節の PSD アプリケーションが示すように，PSD フレームワークでは文書がその文書自身を扱うことができるようなコードを含むのが一般である．このコードを記述するプログラム言語を限定しないことが望ましいため，PSD を実行するためのプラットフォームは XML 文書を処理する XML エディタと PSD に埋め込まれたコードを実行する外部評価系とに切り分けられる．このような構成下では，外部評価系からエディタが保持す

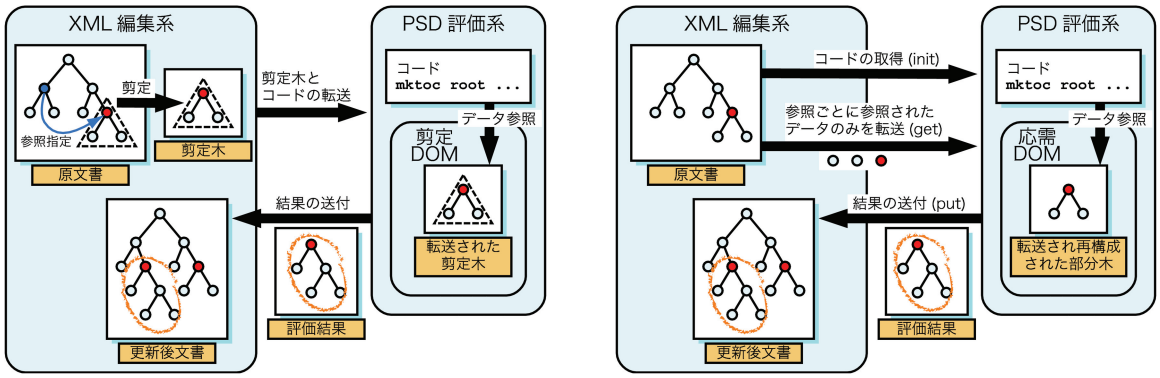

図 3: 剪定木 DOM（左）および応需 DOM（右）

るデータに対して自由に参照・更新できる機構が必要になる．

例えば，TreeCalc の場合には XML 文書としての TreeCalc を編集・表示する XML エディタとして Java で実装された XML 操作環境を用いる一方で，コードの評価系は CDuce といわれる言語で記述されている．このため TreeCalc では，Java の DOM に対して CDuce から自由に参照・更新するための機構が必要となる．単純に DOM オブジェクト全体に相当する XML を生成して外部評価系に渡すという手法では，巨大なデータを扱う場合，特に埋込みコードによる評価が部分的なデータにしか依存しない場合には著しく無駄の多い処理が行われることになる．

この問題を解決するために平成 15 年度にから本年度にかけて，剪定木 DOM（Pruned-Tree DOM）および応需 DOM（By-Need DOM）という二つの PSD プラットフォーム向けの汎用 DOM インタフェースの提案および実装を行った．これらはいずれも外部評価系から DOM データのうち必要な部分のみにアクセスすることができるように設計されており，PSD プラットフォームの性能改善の達成を可能としている（図 3 参照）．すなわち，剪定木 DOM では XML が構造化された文書であることを利用し，コードの評価に必要な部分に関する記述を文書自身に統合し，それを XML エディタ側で解釈することによって不必要なデータのやりとりを回避している．一方，応需 DOM は，外部評価系と XML エディタの間に介在し，計算に必要となった部分のみをその時に応じて DOM より取得し外部評価系に渡すことで，不必要なデータのやりとりを回避するものである．この方式は通信時のオーバーヘッドが増えるが，関数型言語など遅延評価機構を備えた汎用プログラミング言語と組み合わせることで，より効率的な評価を実現することが可能となる．

2.3 双方向計算機構

文書の編集作業，特に XML 文書の編集作業は，直接 XML 文書に手を加えるのではなく，エディタを通じて行なうのが通常である．このエディタでの編集操作は，もともとの XML 文書から一部を抜き出したものに対して行なわれる．この View と呼ばれるエディタ上の表示情報に対して編集が行なわれたとき，変更された結果をもとの文書に反映させなければならない（図 4）．

この問題の典型例として，目次を自動的に生成するようにした文書が挙げられる．目次情報はもともとの XML 文書にはなく，文書内に埋め込まれたコードによって目次情報が生成されるようにするのが普通である．この文書の View に対して，文書本体の章タイトルを変更した場合も目次を変更した場合も同様にもともとの XML 文書に変更が加えられ，目次情報および文書本体の章タイトルがそれぞれ更新されるのが望ましい．

このような表示に基づく編集機能を設計する際に困難なことは，どのような変更もその結果がも

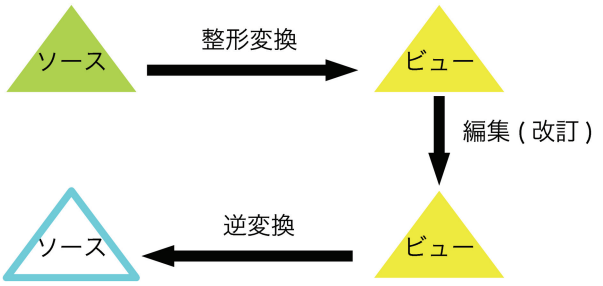

図 4: 双方向編集機構

との文書に反映させることを保証することができ，かつ妥当な変換規則の制約を定めることである．平成 15 年度から本年度にかけて，このような仕様記述に適した双方向に変換可能な言語に関する研究を行い，Inv という双方向 XML エディタのための単射言語を提案した．そして，Inv により，XML 文書のための構造化エディタの構成が可能であることを確認することができた．また，双方向変換を可能とする言語がどのような構成要素を持つのかを明らかにし，XML 編集に利用されている HaXML と呼ばれる特定領域言語（Domain Specific Language）で行ない得る操作はすべて双方向変換言語の構成要素の組合せによって表現可能であることを示した．

3　研究活動

本年度の研究活動としては，上記の研究成果を国際会議等で発表したほか，2005 年 1 月 26 日〜28 日には，横浜市開港記念会館で The Third Workshop on Programmable Structured Documents（第 3 回 PSD ワークショップ）を開催した．国内外から 6 名の関連分野の研究者を招聘し，30 名近い参加者を得て，研究発表・討論による密度の高い研究交流を行った．なお，このワークショップは，本 21 世紀 COE プログラムと，文部科学省リーディングプロジェクト e-Society「高信頼性構造化文書変換技術」，および科学研究費補助金基盤研究 (A)(2)「運算随伴方式による文書情報処理言語の設計とその効果的利用に関する研究」により実施したものである．

本年度発表論文等（抜粋）

論文

- 横山哲郎, 胡振江, 武市正人, “決定論的 2 階パターンとプログラム変換への応用,” **コンピュータソフトウェア**, Vol. 21, No. 5, pp. 71–76 (2004).
- Tetsuo Yokoyama and Zhenjiang Hu and Masato Takeichi, “Deterministic Second-order Patterns,” *Information Processing Letters*, Vol. 89, No. 6, pp. 309–314, (2004).

国際会議論文

- Shin-Cheng Mu, Zhenjiang Hu and Masato Takeichi, “An injective language for reversible computation.” *Mathematics for Programming Construction (MPC2004)*, Stirling, Scotland, UK., July 12–14, *LNCS* 3125, pp. 289–313, 2004.
- Zhenjiang Hu, Shin-Cheng Mu and Masato Takeichi, “A programmable editor for developing structured documents based on bidirectional transformation,” *Partial Evaluation and Semantics-Based Program Manipulation (PEPM'04)*, Verona, Italy, August 24–25, ACM Press, pp. 178–189, 2004.
- Zhenjiang Hu, Kento Emoto, Shin-Cheng Mu, Masato Takeichi, “Bidirectionalizing Tree Transformations,” *International Workshop on New Approaches to Software Construction (WNASC 2004)*, The University of Tokyo, Komaba, Tokyo, Japan, September 13–14, 2004. pp.3–22.
- Shin-Cheng Mu, Zhenjiang Hu and Masato Takeichi, “An algebraic approach to bidirectional updating problem,” *The 2nd Asian Symposium on Programming Languages and Systems (APLAS2004)*, Taipei, Taiwan, Nov. 4–6, *LNCS* 3302, pp. 2–20, 2004.

特許出願

- 胡振江，穆信成，武市正人．「構造化文書作成方法及び装置及びプログラム」(特願 2004-242265, 平成 16 年 8 月 23 日)