D-011

XML 文書コンパクト化 (CSV 圧縮)の変換仕様作成用エディタの開発

Development of an Editor for Specifications to Compact XML Documents with CSV Format

吉田 茂† Shigeru Yoshida　中島 哲† Satoshi Nakashima　小田切 淳一† Junichi Odagiri　伊藤 秀一†† Shuichi Itoh

1．まえがき

XML は，柔軟で拡張性が高いデータ表現形式であるが，データ処理負荷が高く，メモリを大量に消費する課題がある。レコード形式の大容量 XML 文書の負荷を軽減するため，複数個の要素をレコード単位に CSV(Comma Separated Values)形式でまとめる方法「XML CSV 圧縮」を，先に発表した[1]。しかしながら，この方法は，事前に圧縮の仕方を指定する「変換仕様」を作成する必要があり，階層構造が複雑な文書や，多数の要素を持つ文書で，作成に手間がかかった。これを解決するため，変換仕様を簡単に作成できる支援ツールを開発し，実用的にしたので，報告する。

2．XML CSV 圧縮について

(1) XML CSV 圧縮の概要

標準 API の DOM 処理[2]は XML 文書の全要素をメモリ上に展開するため負荷が重くなる。しかし，応用ソフトでは全種の要素をレコード横断のデータ処理の対象にすることは少ないので，本技術では，冗長な情報を一つにまとめて要素数を減らすようにする。応用ソフトでレコード横断的に操作対象とする要素（生要素)をそのままにして，それ以外の複数の要素（CSV 化要素)の要素内容を CSV 形式で一つの要素（CSV 要素)にまとめた XML 文書に変換する(図１)。CSV 化要素は，変換で付加されるヘッダを見て，要素名に対応付けて格納要素内容を取り出せる。本技術により，DOM 処理や XSL 変換 [3]において，CSV 化で減らした要素の割合にほぼ比例して，メモリ使用量，メモリ展開時間を削減することができる。

(2) 課題

CSV 圧縮は XSL 変換として実行される。レコード中の要素名を出現順に列挙して，CSV 化要素を指定する「変換仕様」（図２)の XML 文書をユーザが作成すれば，その XSL シートは自動的に生成するようにした。しかし，レコード中に数十～数百個の要素や，深い階層を持つ XML 文書の変換仕様をテキスト・エディタで作成するのは，手間が掛かり，ミスも生じ易いという課題があった。これを解決するために，支援ツールとして，変換仕様を GUI を用いて簡単に設計できる Java ソフトのエディタを開発した。

3．開発技術

(1) GUI エディタの使用と CSV 圧縮変換のフロー（図３）

処理対象の XML 文書から抽出した DTD(文書型定義)を，本エディタに読込んで，CSV 圧縮の仕方を指定し，変換仕様を作成・出力する。対象 XML 文書から DTD の抽出はフリーソフトを用いて行える[4]。作成した変換仕様から CSV 圧縮用ソフトを用いて，圧縮/復元用 XSL シートを自動生

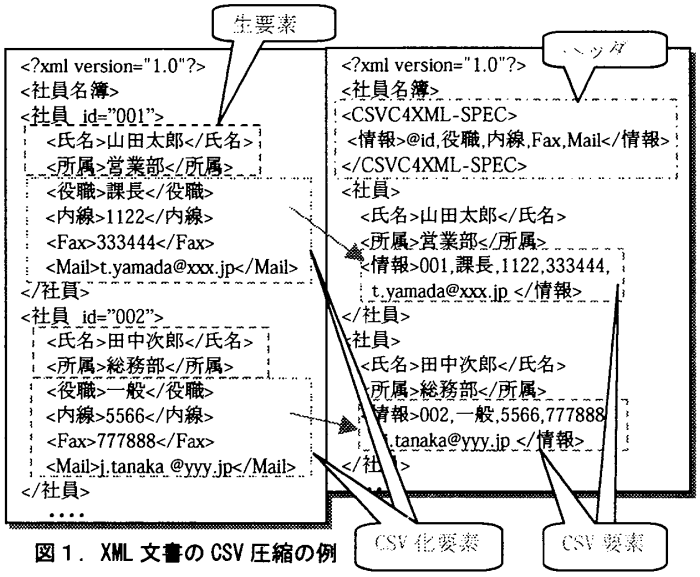

図１．XML 文書の CSV 圧縮の例

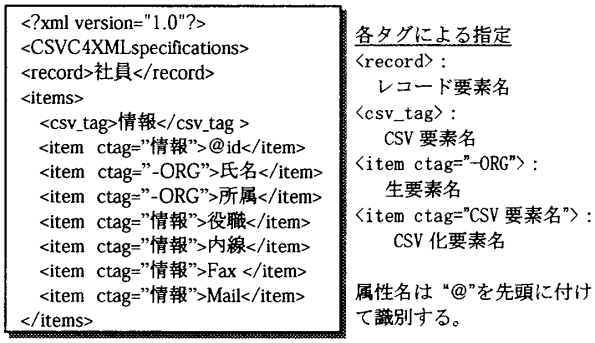

図２．変換仕様ＸＭＬ文書の例

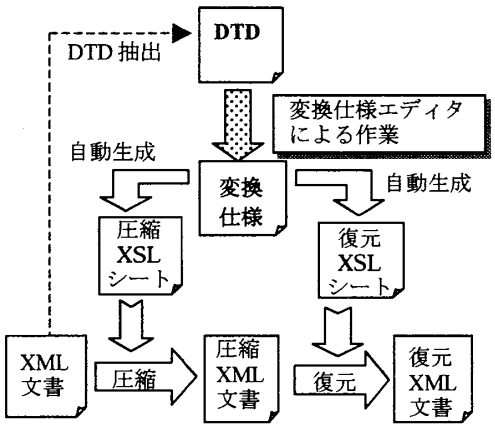

図３．変換仕様エディタの作業を含む CSV 圧縮・復元のフロー

† （株）富士通研究所 ITS 研究センター
†† 電気通信大学大学院 情報システム学研究科

成する。次に，圧縮用 XSL シートを用いて，事前に圧縮 XML 文書を作成し，それを応用ソフトでデータ処理する。最後に，処理結果の XML 文書を復元用 XSL シートを用いて，元の形式の XML 文書に戻す。

(2) 変換仕様エディタの機能（図５）

(a)処理対象の XML 文書の DTD（図４）を読込み，データ構造と全要素名を画面に表示する。

(b)画面上の要素を選択し，操作パネルのボタンをクリックして，レコード要素(Record Elem)，生要素(Original Item)，CSV 化要素(CSV Item)を指定する。指定された要素は画面上で色分けして区別される。

(c)指定結果を変換仕様として出力する。このとき結果の変換仕様の作業用ファイルも出力しておくことで，それを再度読込んで，追加修正・変更が行える。

(3) 複雑な XML 文書への対応

次のような指定により種々の XML 文書に対応できるようした。

(a)任意の階層

DTD を読み込むとエディタ画面上に，階層がツリー表示される。末端の葉のノードを選択し，CSV 化の指定を行う。

(b)非定型要素（出現に有無があるもの）

DTD を読み取り，図６のように非定型要素を区別して表示する。ユーザは定型/非定型を意識せずに指定できる。出力される変換仕様から作られる圧縮/復元用 XSL シートには，非定型要素の場合，復元時に要素内容の有無を検査して空要素を削る操作が組み込まれる。

(c)同名要素の繰返し

同名要素は対象のノードに最大繰返し数を指定して，最大個数分の要素を表示させ，CSV 化の指定を行うことができる。この表示された複数個の要素は上記(b)の非定型要素とみられ，復元時に要素内容があるものだけが再現される。

(d)レコードのネスト（子レコード）

レコード中に複数要素を持つレコードが存在する場合，複数個のレコード要素を指定できるようにした。出力される変換仕様から作られる圧縮/復元用 XSL シートには，指定されたレコード要素ごとにその操作を template として組み込むようにした。

４．機能の評価

XML コンソーシアム [5]が策定した標準フォーマットである ContactXML, TravelXML 他を用いて，変換仕様エディタの機能を確認した。１時間程度かかっていた変換仕様の作成作業が，短時間で行えるようになった。

５．むすび

本技術の特長は，応用ソフトでレコード横断的に操作する要素の種類が予め分かれば XML 文書を必要最小限の DOM 木で扱えること，XSL 変換として実行するため種々の言語で利用できること，プログラムレスで簡単に実行できること，である。本技術は，大容量 XML 文書を扱う以外に，CSV 圧縮形式を経由しての XML 文書の形式変換にも，有用と考える。本技術の評価版ソフトは富士通研究所のサイトから配布している[6]。試用して頂き，ご意見，ご感想を頂ければ有り難い。

参考文献

[1] 吉田他「XML 文書の事前形式変換によるデータ処理性能改善の検討」FIT2002 予稿集 D-29, 2002.9.27

[2] DOM http://www.w3.org/DOM

[3] XSLT http://www.w3.org/TR/xslt

[4] DTDGenerator A tool for generate XML DTDs http://saxon.sourceforge.net/dtdgen.html

[5] XML コンソーシアム http://www.xmlconsortium.org/

[6] 富士通研究所 フリーソフトの配布サイト http://www.labs.fujitsu.com/jp/freesoft/csvc4xml/index.htm


```
<!ELEMENT 社員名簿 ( 社員+ ) >
<!ELEMENT 社員 ( 氏名, 所属, 役職, 内線, Fax, Mail ) >
<!ATTLIST 社員 id NMTOKEN #REQUIRED >
<!ELEMENT 氏名 ( #PCDATA ) >
<!ELEMENT 所属 ( #PCDATA ) >
<!ELEMENT 役職 ( #PCDATA ) >
<!ELEMENT 内線 ( #PCDATA ) >
<!ELEMENT Fax ( #PCDATA ) >
<!ELEMENT Mail ( #PCDATA ) >
```

図４．図１の XML 文書の DTD (Document Type Definition)

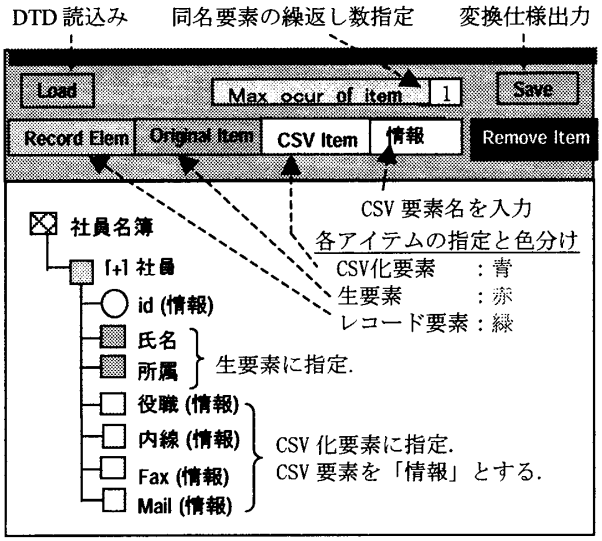

図５．DTD を入力し指定後の変換仕様エディタ画面の例

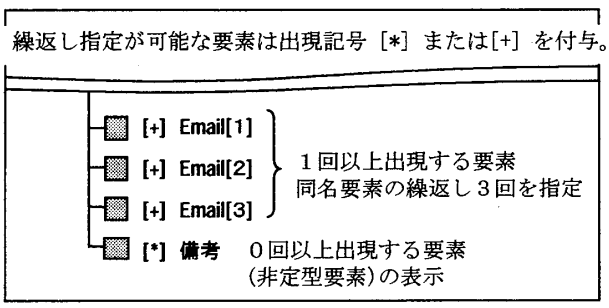

図６．エディタ画面の同名要素繰返し，非定型要素の表示例