多様な利用者環境で多言語文書表示を可能にするXMLブラウザ

阪口 哲男, 永森 光晴, 杉本 重雄, 田畑 孝一
図書館情報大学
〒305-8550 茨城県つくば市春日 1-2
E-mail: {saka, nagamori, sugimoto, tabata}@ulis.ac.jp

概要

現在、WWW において HTML に代わる新たなマークアップ言語 XML が注目されており、これを実際に利用するためのツールが提供されつつある。XML は多言語文書のために Unicode を採用している。しかしながら、実際にはインストール済みの文字フォントを始めとする利用者環境に影響を受け、XML 文書を表示できないことも多い。本研究では、著者らが開発した MHTML Browser 開発で得られた知見に基づき、多言語 XML 文書を多種多様にわたる利用者環境において表示することを目的とする。今回はシステムの構想及び開発状況について発表する。

キーワード

多言語文書, XML, 文字フォント, MHTML

A XML Browser of Multilingual Documents for Users of Various Computing Environments

Tetsuo Sakaguchi, Mitsuharu Nagamori, Shigeo Sugimoto, Koichi Tabata
University of Library and Information Science
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, JAPAN
E-mail: {saka, nagamori, sugimoto, tabata}@ulis.ac.jp

Abstract

XML is a new markup language defined by the WWW Consortium. Many software tools for processing XML documents are provided freely or commercially and used to develop applications based on XML. The Unicode character set is used for writing multilingual XML documents. However, not all computing environments are able to display multilingual documents because they do not have enough character fonts or have some other problems. This paper describes a XML browser of multilingual documents for users of various computing environments. It discusses problems of existing XML browsing tools and design and implementation of the browser.

Keywords

multilingual documents, XML, character font, MHTML

1. はじめに

World Wide Web (WWW) ではその記述言語として HTML (Hypertext Markup Language) が用いられている。しかしながら HTML には制約も多いため、より自由度の高いマークアップ言語である XML (Extensible Markup Language) が WWW コンソーシアム (W3C) によって提案されており、様々な分野で利用されつつある。

XML では多言語文書の記述を可能にするために、文字セットとして Unicode を採用している。Unicode には様々な言語のための文字が含まれているが、現状ではあらゆる環境でそれらのすべての文字を正しく表示できるとは限らない。

本研究では、XML で記述された多言語文書を様々な利用者環境で表示可能にするブラウザの開発を目的とする。著者らは以前 HTML で記述された多言語文書を表示可能にする MHTML Browser を開発しており、その開発過程で得た知見に基づいて本ブラウザの開発を進める。

2. XML 文書の表示と多言語対応

XML では Unicode が採用されているため、多言語文書の記述は規格上では可能となっている。Unicode に対しては様々な問題点の指摘もされているが、Java のような Unicode をサポートしたプログラミング言語も普及しており、Unicode で書かれた多言語 XML 文書を処理する環境が整いつつある。一方、XML 文書の表示については、必要な文字フォントが備わっていなかったり、禁則処理などの各言語に依存した処理が行われなかったりして、正しく表示されない場合もまだ多い。

本節では XML 文書の表示における多言語対応の現状について検討する。XML 文書を表示する環境としては様々なものがあり、クライアント側で XML 及びスタイルシートの処理を行って表示する方式と、サーバ側で XML とスタイルシートの処理を行い HTML や PDF などクライアント側で表示可能な形式に変換する方式の 2 つの方式が存在する。前者の例としては、XML 対応の WWW ブラウザとして最も普及していると思われる Microsoft Internet Explorer (MS-IE) を、後者の例としては Apache XML Project[1] の一環として開発されている Cocoon を取り上げる。

MS-IE はバージョン 4 から XML 対応が進められており、バージョン 5 では Cascading Style Sheets Level 2 (CSS2)[2] によって XML 文書を整形して表示する機能を備えている。MS-IE では文書を表示する際にオペレーティングシステム (OS) に組み込まれている文字フォントを主として用いる。そのため、多言語文書を表示する際にはあらかじめ各言語対応の文字セットを備えたフォントを導入しておく必要がある。例えば、MS-Windows 2000 では様々な言語対応のフォントを選択して追加導入することが可能である。また、CSS2 ではフォントを WWW サーバからダウンロードして使用することを指定できる。そのため、必要なフォントを WWW サーバ上に準備し、そのダウンロードを指定したスタイルシートを使用することで、OS に組み込まれていないフォントを使って表示することが可能である。MS-IE では文書の記述方向として左右両方向に対応している他、バージョン 5.5 では縦書きにも対応している。その他の各国語対応機能としては日本語の禁則処理などにも対応している。

Cocoon は WWW サーバ上で稼働する。サーバ上に XML 文書とともに XSL (Extensible Stylesheet Language) で記述されたスタイルシートを格納しておき、XML 文書をスタイルシートによって整形・変換したものを WWW ブラウザに送って表示する。変換後の形式は一般的な WWW ブラウザによって理解できる必要があるため、HTML や PDF 形式が選ばれる。その結果、多言語表示に関しては WWW ブラウザとそれが稼働している OS などに依存することになる。また HTML については CSS2 などによりフォントのダウンロード指定をすることが、PDF についてはフォントの埋め込みが可能となっているので、OS に組み込まれていないフォントを使うことも可能である。

以上のように、多言語 XML 文書の表示については通常はブラウザが稼働している環境、特に OS が備えている文字フォントの種類に依存する。また、OS が備えていない文字フォントに関して WWW サーバなどからダウンロードするように指定することは可能であるが、実際にどのフォントをダウンロードすれば良いかは個々の OS の種類や設定に依存する。また、OS やブラウザによって使用できるフォントのデータ形式に違いがあるため、それらすべてに対応するように WWW サーバ上にフォントを準備するのは難しい。Cocoon で PDF に変換する場合は、フォントのデータ形式の問題は回避できるが、一般に PDF データの表示には WWW ブラウザとは別に PDF 表示用のソフトウェアが要求される。

以上のように、XML の規格上は多言語文書を扱うことが出来るが、正しく表示できるかどうかは、実際の利用者の環境に大きく依存することになる。現状では各種言語のフォントをすべて追加導入した MS-Windows 2000 と MS-IE5.5 の組合せが最良であるが、そのように利用者環境の統一をインターネット全体で行うことは事実上不可能である。

近年、インターネットの利用者環境はパーソナルコンピュータ (PC) やワークステーション (WS) のみではなく、携帯情報端末や携帯電話なども加わり多様化している。そのような多様な利用者環境においても様々な言語で書かれた文書を正しく表示することが求められる。

3. MHTML Browser

WWW では様々な言語で書かれた HTML 文書が提供されている。MHTML (Multilingual HTML) Browser は、文字フォントがない環境においてそのような多言語 HTML 文書を表示するものである [3][4]。MHTML Browser は図 1 のような構成をとっており、Gateway によって HTML 文書にその表示に必要な分だけの文字フォントを付加してクライアントに送る。クライアント上ではそのフォントを用いて HTML 文書の表示を行う。表示機能は Java Applet によって実現しているので、Java Applet を実行できる WWW ブラウザが備わっていれば、どんな環境でも多言語 HTML 文書を表示することが可能となる。

＝＝＝＝＝ 図 1 MHTML Browser の構成 ＝＝＝＝＝

この MHTML Browser は利用者が URL を指定して目的のページを見ることが出来るほか、「日本昔話の多言語ディジタル文庫」[5] のように文書提供者の側であらかじめ MHTML Browser を組み込んで提供することも可能である。この組み込み機能を用いて、現在 Dublin Core メタデータの様々な言語で記述された参照記述の提供も行っている [6]。

このように MHTML Browser によって様々な言語の WWW ページを見ることが可能になるが、以下のような問題点も残されている。

(1) 禁則処理やワードラッピング処理の欠如

(2) 左から右への横書きのみの対応であり、アラビア語などの右から左に記述するような言語に未対応であること

現在の MHTML Browser では Gateway は文字フォントを付加する処理に留まっており、HTML タグの解釈と表示レイアウト処理はすべてクライアント上にある Viewer が行っている。この Viewer においては言語依存の処理をしない構成をとっており、そのままではこれら 2 点の問題を解決するのは困難である。

また、システム実現上の問題としても性能の低さや、操作性、対応している HTML タグの種類の少なさ、組み込み機能における URL や言語の指定方法などにも問題が残されている。

4. 多様な利用者環境における多言語 XML ブラウザ

前述のように MHTML Browser にはいくつかの問題点が残されているものの、文字フォントなどの利用者環境に依存することなく多言語文書を表示可能としている点は情報提供する際の大きな利点である。著者らは現在メタデータのレジストリの開発を行っており、そこで様々な言語で記述されたメタデータの表示のためにも MHTML Browser 相当の機能が必要であると考えている [7]。また、メタデータレジストリでは、メタデータの記述形式として RDF (Resource Description Framework) を前提としているため、表示機能としても XML に対応していることが望まれる。XML は HTML に比べて文書構造をより自由に表現することが出来るため、今後の情報流通の際のデータ形式として有望視されている。

そこで本研究では XML で記述された多言語文書を対象としたブラウザの開発を行う。XML を採用することによって、HTML を XML 規格に基づいて再定義した XHTML 規格への対応も可能であり、また何らかの前処理を施すことによって現存する多くの HTML 文書の表示も可能であると考えられる。

本稿冒頭で述べたように XML では Unicode が採用されており、データ形式として多言語文書の記述は可能である。しかしながら、2 節で述べたようにその表示についてはまだまだ利用者環境に大きく依存し、どこでも多言語文書を正しく表示できるという状態には至っていない。また、XML 文書を表示するには文字フォントを備えているだけではなく、XSL の処理機構なども備えることになるため、ある程度の処理能力を備えた環境が求められる。そのような要求を極力減らし、より多くの利用者環境で多言語 XML 文書の表示を可能にするブラウザの開発を行う。

携帯電話などの処理能力が低い環境を想定して考えられたブラウザの規格として WAP (Wireless Application Protocol) がある [8]。WAP では MHTML Browser と同様に Gateway を設置し、その Gateway である程度の処理を行った結果を端末に送ることで端末装置の負荷と通信データ量を軽減することを狙っている。今回開発するシステムでは、この WAP を参考にして、MHTML Browser のシステム構成を受け継ぎつつそれぞれの構成要素の役割分担を以下のように見直した。

(a) Gateway では、文字フォントを付与するだけでなく、XML タグの解釈とスタイルシートの処理を行って、表示に必要なレイアウト情報を生成する。そのレイアウト情報には禁則処理やワードラッピング、記述方向に関するものも含める。

(b) Viewer では、Gateway から渡されたレイアウト情報に基づき、付与された文字フォントを用いて実際の画面に合わせた表示を行う。

Gateway 側でより多くの処理を行うことによって Viewer 側の処理が軽減されるので、MHTML Browser に比べてクライアントに対する負荷の軽減が期待できる。また、禁則処理など言語に依存した処理を Gateway 側で行うことにより、Viewer の変更なしに対応言語を拡張することも可能となる。

開発については、Viewer だけでなく、Gateway 機能も Java を用いて開発することにより、Gateway の移植性も高める。XML のタグ処理及びスタイルシート処理に関しては既存のライブラリなどを利用することを考えている。なお、Unicode 以外の文字コード系への対応や HTML への対応については前処理系を準備することで対応する。

5. おわりに

本稿では多様な利用者環境で多言語 XML 文書を表示するブラウザのシステムの構想を中心に述べた。現在、システムの開発は 2001 年 3 月頃に最初の版を公開することを目標に進めている。

参考文献

[1] The Apache XML Project. http://xml.apache.org/

[2] WWW Consortium. Cascading Style Sheets, level 2 - CSS2 Specification. W3C Recommendation 12-May-1998. 1998. (URL: http://www.w3.org/TR/1998/REC-CSS2-19980512)

[3] 前田亮, Myriam Dartois, 太田純, 藤田岳久, 阪口哲男, 杉本重雄, 田畑孝一. クライアントにフォントを必要としない多言語 HTML 文書ブラウジングシステム. 情報処理学会論文誌, Vol.39, No.3, pp.802-809. 1998.

[4] Akira Maeda, Myriam Dartois, Takehisa Fujita, Tetsuo Sakaguchi, Shigeo Sugimoto, Koichi Tabata. Viewing Multilingual Documents on Your Local Web Browser. Communications of the ACM, Vol.41, No.4, pp.64-65. 1998.

[5] 日本昔話の多言語ディジタル文庫. http://www.DL.ulis.ac.jp/oldtales/

[6] 永森光晴, Thomas Baker, 阪口哲男, 杉本重雄, 田畑孝一. Dublin Core Metadata Element Set における多言語への対応. 情報処理学会研究報告 (情報学基礎研究会 99-FI-56), Vol.99, No.102, pp.1–8. 1999.

[7] 永森光晴, Thomas Baker, 阪口哲男, 杉本重雄, 田畑孝一. RDF Schema に基づくメタデータレジストリ. 情報処理学会研究報告 (情報学基礎研究会 2000-FI-60), Vol.2000, No.91, pp.1–8, 2000.

[8] Wireless Application Procotol Forum. WAP Architecture, Version 30-Apr-1998. 20p. 1998. (http://www.wapforum. より入手可能)

その他、XML, XSL 関係規格は W3C (http://www.w3.org/) より入手可能。