組み込みフォントを必要としないWWWのための多言語ブラウザ

前田亮，藤田岳久，リー スエイチュー，阪口哲男，杉本重雄，田畑孝一
図書館情報大学
〒305 茨城県つくば市春日 1-2
Tel: 0298-52-0511 Fax: 0298-52-4326
E-mail: {maeda, take, lee, saka, sugimoto, tabata}@ulis.ac.jp

概要

World-Wide Web(以下 WWW) は様々な国で使われ，様々な言語のドキュメントが提供されている．Mosaic や Netscape などの WWW ブラウザでは，フォントを用意すれば様々な言語で書かれたドキュメントを表示することが可能であるが，現状では多言語で書かれたドキュメントを表示することはできない．また，ユーザが外国語のドキュメントを読むために必要なフォントをすべて用意するのは現実的でないと思われる．

本稿では，WWW 環境において多言語を表示するために開発した簡易ブラウザ，およびそれを実現するための記述方式である MHTML について述べる．このブラウザと記述方式を用いることによって，クライアント側の組み込みフォントを用いずに容易に多言語の HTML ドキュメントを表示することが可能となる．

キーワード

インターネット，WWW，HTML，多言語ブラウザ，国際化

A Multilingual Browser for WWW without Preloaded Fonts

Akira Maeda, Takehisa Fujita, Lee Swee Choo, Tetsuo Sakaguchi, Shigeo Sugimoto, Koichi Tabata
University of Library and Information Science
1-2, Kasuga, Tsukuba, Ibaraki, 305, JAPAN
Phone: +81-298-52-0511 Fax: +81-298-52-4326
E-mail: {maeda, take, lee, saka, sugimoto, tabata}@ulis.ac.jp

Abstract

The World-Wide Web (WWW) provides us with documents produced in various countries. Conventional WWW browsers such as Mosaic and Netscape have facilities to browse a document written not only in English but also in other languages, e.g., Japanese, Korean, Chinese and French. However, those browsers are not useful for a document which is written in more than one language, e.g., Rosetta stone. In addition, it is not feasible for users to make all fonts required to browse foreign documents available in their machines.

This paper describes a light-weight browser for multilingual documents in the WWW environment. The browser displays HTML text written in a foreign language but requires no preloaded fonts for that language. This paper presents a document description scheme for multilingual documents called MHTML. It also shows a prototype MHTML browser and its performance.

Keywords

Internet, WWW, HTML, multilingual browser, internationalization

1. はじめに

現在 WWW は世界中の様々な国で広く使われているが，真の国際的なドキュメント流通環境を実現しているとは言い難い面がある．これには大きく分けて 2 つの理由があろう．1 つは，現在の HTML の仕様 [1] では文字セットとして ASCII および ISO-8859-1 (西ヨーロッパの主要言語) の使用しか規定しておらず，それ以外の文字セットの使用に関しては独自に取り決めを行なう必要があることである．もう 1 つは，外国語のドキュメントを読む時に必要なすべてのフォントをローカルな計算機上に事前に揃えておくのは現実的でないということがある．自国語以外には ASCII や ISO-8859-1 などのフォントしか持っていないことが多いのが現状であろう．そのため，例えば日本で中国語のドキュメントを読もうとしてもフォントがなくて読むことができない，といったことが起こる．ディジタル図書館という観点からみても，蓄積・提供するドキュメントの言語を制限することはふさわしくなく，この点からもこの問題を解決する必要があると思われる．

MHTML のドキュメントは，HTML で記述された文章とそれを表示するのに必要なフォントの集合から構成される．MHTML のブラウザを用いれば，クライアント側に組み込みフォントを用意しておく必要がないため，外国の WWW サーバにアクセスするために必要なフォントを探してきたり組み込んだりする手間がなくなる．図 1 に MHTML による方法の概要を示す．

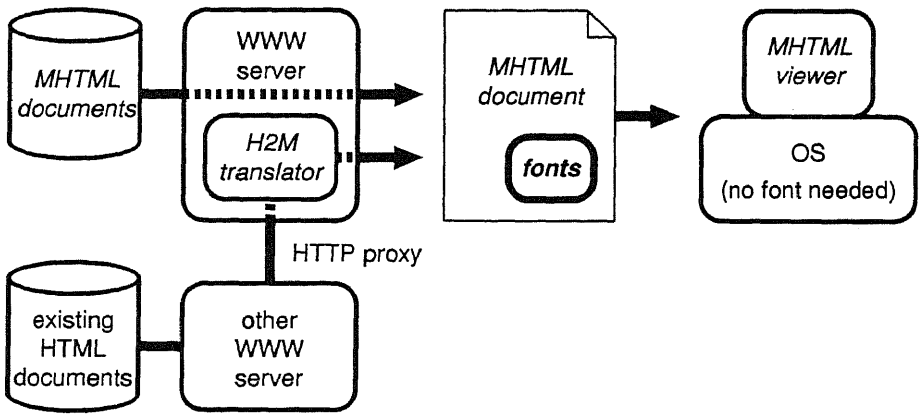

図 1

2. MHTML の構成

MHTML ドキュメントは通常の HTML ドキュメントから生成する．MHTML ドキュメントはヘッダ部，フォント部，テキスト部の 3 つの部分から構成される．ヘッダには，フォント部とテキスト部の先頭へのオフセットが格納される．フォントは，元となる各言語ごとのフォントファイルから文字ごとに抽出される．テキスト部分は，ASCII コードセットおよびドキュメント依存のコードセットの 2 つのコードセットによって符号化される．ドキュメント依存のコードセットは，そのドキュメントに含まれる文字のみから構成される．つまり，変換の手順はまず元となる HTML ドキュメントを読み，そのドキュメントに含まれる文字の集合を調べ，それぞれの文字にコードを割り当て，その文字のフォントをフォントファイルから抽出し，それを元に MHTML ファイルを出力する．現在の実装では，HTML のタグのみに ASCII コードセットを用い，それ以外の部分はすべてドキュメント依存のコードセットを用いている．

3. MHTML ブラウザの試作

上で述べたMHTMLに基づくブラウザを試作した．これにはHTMLからMHTMLへの変換機能(サーバ部分)とMHTMLドキュメントを表示するビューア機能(クライアント部分)が含まれる．ビューア機能はMosaicの外部ビューアとして起動することができる．Mosaicがファイルの拡張子が".mhtml"であるMHTMLドキュメントを受け取るとビューアが起動され，そのドキュメントが表示される．

ビューアではクライアント側にあるフォントを一切用いずに，MHTMLドキュメントに含まれるフォントのみを用いて表示を行なう．図2に英語，日本語および韓国語を含んだドキュメントを表示している様子を示す．アンカー(ハイパテキストリンク)の部分はMosaicなどと同様に下線が引かれている．アンカー部分をクリックすると，ビューアはMosaicに対して指定されたファイルを表示するように要求する．

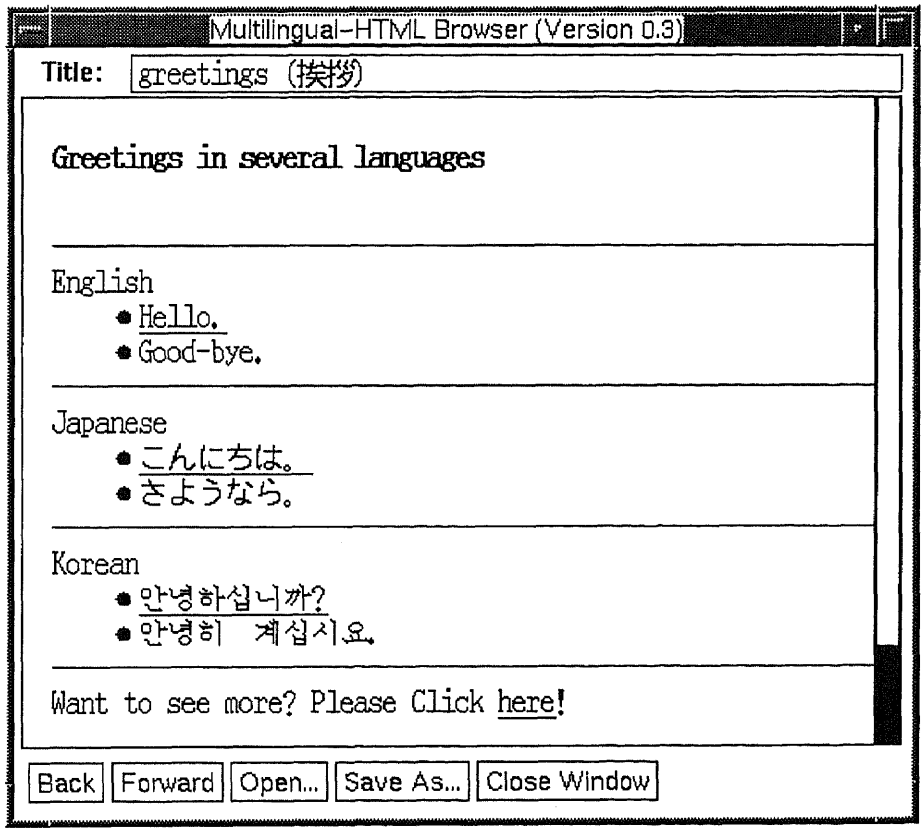

図2

4. 評価

4.1 ドキュメントの大きさ

MHTMLのドキュメントはテキスト部分に加えフォントを含んでいるため，元のHTMLドキュメントよりも大きくなる．幾つかのドキュメントについて調べたところ，MHTMLと元のHTMLの大きさ(バイ

ト数) の比 (MHTML/HTML) は約 1.7 から 5.0 であり，ドキュメントが大きくなるに従って比は小さくなる傾向がある．

元のテキストに含まれる全文字数を L とし，互いに異なる文字数を C とすると，MHTML/HTML は次の式で与えられる．

$$\frac{\mathrm{MHTML}}{\mathrm{HTML}} = \frac{2L + C \times \alpha}{2L} \quad (\alpha = 32)$$
$$= 1 + \frac{C}{L} \times \frac{\alpha}{2}$$

ここで α は 16 × 16 ドットの日本語文字 1 文字に必要なバイト数である．ただし実際の MHTML では，8 × 16 ドットの英文字や，フォントを含まないタグも含まれるため，この式とは異なる．元のテキストが長くなるに従って C/L が小さくなるのは明らかである．英語の文章の場合，C の上限は ASCII の全文字数であるが，日本語の文章の場合は実際の C の上限は JIS の全文字数よりかなり小さくなると考えられる．例えば，5,368 文字と 16,615 文字の 2 つの日本語のドキュメントについて調べたところ，C/L はそれぞれ 0.11 と 0.05 であった．

4.2 他の方法との比較

WWW 上で，組み込みフォントを使用しないで多言語の HTML ドキュメントの表示を実現する方法として，他にも次のようなものが考えられる．

- イメージマップ: HTML ドキュメントをイメージ (クリッカブル) マップに変換してクライアントに送る．
- 文字毎のインラインイメージ: HTML ドキュメントに含まれる各文字をインラインイメージとしてクライアントに送る [2][3].

MHTML による方法との比較のため，1 つの HTML ドキュメントをこれらの方法により変換したものを用意し，実験を行なった．この結果を表 1 に示す．

表 1

	MHTML	Image map	Inline image
Total Data Size (Kbytes)	6.3	16.7	33.5
Display Speed (sec.)	1.5	6.0	21.0

5. おわりに

MHTML およびブラウザの今後の課題として，次のようなものが挙げられる．

- 様々な大きさのフォント，およびプロポーショナルなフォントへの対応．

- データ量を減らすため，MHTMLドキュメントの圧縮.
- MHTMLビューアにおける入力機能(カットアンドペースト，formによる入力など)の実現.
- Proxy機能[4]による他のサーバのHTMLドキュメントのMHTMLへの自動変換.
- 多言語ドキュメントの編集機能.

多言語に対応したブラウザは今後の“wall-less”なディジタル図書館にとって必須のツールとなるであろう．しかし，インターネット上での多言語環境はまだ一般的でないのが現状である．MHTMLによる方法は単純ではあるが，多言語環境を実現するための現実的な方法として有効であると思われる．

参考文献

[1] Berners-Lee, T., Connolly, D., HyperText Markup Language – 2.0, Internet Draft, 71p, 1995.

[2] Shopov, V., Character to Inline Image Conversion Library (CIILIB), 1995.
<URL:http://baka.aubg.bg/readme.cii>

[3] Sato, Y., What is the DeleGate ?, 1995.
<URL:http://www.etl.go.jp:8080/etl/People/ysato@etl.go.jp/DeleGate/>

[4] Luotonen, A., Altis, K., World-Wide Web Proxies, WWW'94 Conference, 8p, 1994.
<URL:http://www.w3.org/hypertext/WWW/Proxies/>