インターネットにおける効率的なフォントデータの配送方式

4Aa－10

櫻井英樹　山本晋一郎　濱口毅　阿草清滋

名古屋大学工学部

1　はじめに

インターネットの発達にともない世界中いたるところから自分のコンピュータ環境を扱えるようになっている。ところが，日本語データへのアクセスに関しては，利用環境に日本語環境がない場合にはコードとしては自在に入手できても，画面上に意味ある字形として見られない場合が多い。これはわが国に限らず英語圏以外の国々の共通の問題である。

今後のネットワークの進展にともない，アプリケーションには多種類のコンピュータ，多様な言語への対応が求められるが，すべての言語のすべてのフォントをあらかじめ持つことは容量などの点から現実的ではない。そのため，オン・デマンドでフォントデータをやり取りできることが求められる。

すでに X ウィンドウシステムなど，フォントサービス機能を持つシステムは出てきているが，広くインターネット上で利用できるようなものはまだ現れていないのが現状である。そこで著者らはフォントデータの効率の良い配送方式を提案する。本方式では，日本語フォントなどのデータ量が多いものを扱う場合，使用頻度の高いデータから，徐々に配送することにより，効率を向上させる。ここでいう効率的とは，通常のフォントファイルをそのまま配送する場合では，クライアントはフォントデータが最後まで送られてくるまで待っていなければならないが，本方式を用いれば，最後まで待つ必要はなく，続きの処理を先に実行することができるということである。

2　方針

次の二点をポイントに考えている。

- インターネットにつながっているマシンには UNIX や X ウィンドウシステムが載っているという前提は今や成り立たず，MS ウィンドウズやマッキントッシュその他である場合も増えてきている。フォントサービスは，これらの環境に依存しない，汎用性のあるものでなければならない。
- 英語やドイツ語など，すべてのアルファベットが256文字以内に収まってしまうような言語のフォントについては問題ないが，日本語，中国語など，数千文字も必要になるようなフォントを取り扱う場合，配送にかかるオーバーヘッドが問題となる。我々が日本語の文章を書く場合，通常用いる漢字は，かなり限られたものになっていると考えられる。そこで，日本語フォントなどのデータ量が大きいものを扱う場合，ひらがな，カタカナ，通常使われる漢字など，使用頻度の高いデータを先に配送し，残りはバックグラウンドで徐々に送るようにすれば，第一応答を早くすることができると考えられる。

3　研究の内容

今回は，日本語フォントについて実験を行なう。

まず，漢字の重要度の決め方についてであるが，実際コンピュータ上で用いられるテキストデータ中の，各漢字の出現頻度から重要度を求めることにした。それには大量のテキストデータが必要となるが，ここではネットニュースのニュースプールに蓄えられている大量の記事から，漢字の出現頻度を計測した。ニュースプールのデータに対して実験を行なった理由は，それが手軽に入手できる大量のテキストデータである点と，様々なカテゴリの記事があるので，偏りの少ないデータであると考えられる点からである。

この結果，全記事中に現れた漢字の総数は，延べ6,768,511個であった。また，漢字の種類は，4,779種類であった。この中でももっとも出現頻度が高かったのは，“大”の字で，85,156回現れた。この結果の主要部分をグラフに表すと，図1が得られた。ここで，横軸は文書中に現れた漢字の番号で，度数が大きい順に並びかえてある。縦軸は度数である。

このデータから，日本語フォントのビットマップデータをグループ分けする。フォントサーバは，この日本語フォントが要求されたとき，グループ分けされたデータの中から，一番重要度が高いグループをまず配送する。フォントサーバを利用するアプリケーションプログラムをクライアントと呼ぶことにすると，クライアントは，自分の必要とする漢字が，このグループの中にすべて含まれていた場合，それ以降のデータ

Efficient Transferring Font Data through Internet
Hideki Sakurai, Shinichirou Yamamoto, Takeshi Hamaguchi, Kiyoshi Agusa
Nagoya University

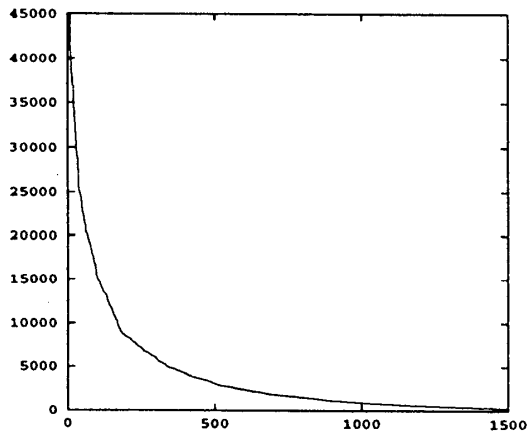

図 1: 度数分布グラフ

が送られてくるのを待たずに，続きの処理 (表示など) を実行する。また，その旨をフォントサーバに通知する。フォントサーバは，残りのグループのデータについても順番に配送する。クライアントは，それらのデータをバックグラウンドで受けとるようにする (図 2)。

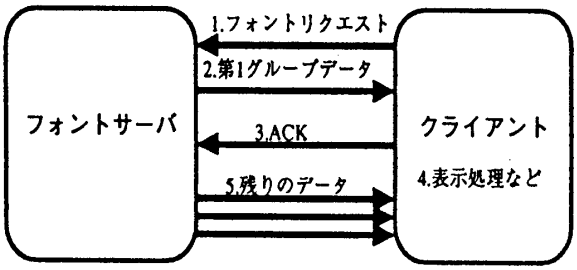

図 2: 理想的な場合

第一のグループ中に，クライアントが必要とする漢字がすべて含まれていなかった場合，クライアントは，足りない漢字についての送信要求をフォントサーバに返す。それを受けてフォントサーバは，不足分のデータをまとめてクライアントに送信する。後の処理は上と同様である (図 3)。

この結果，一度にフォントすべてを読み込む方式に比べて，第一応答が早くなる。これは，インターネットを介して遠く離れたマシン間でフォントをやり取りするような場合，より有効になると考えられる。

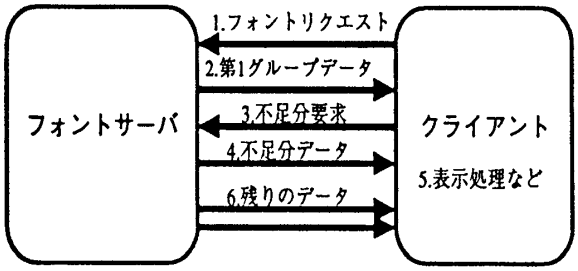

図 3: 足りなかった場合

4 今後の課題

今回は単純に，ニューススプール中の全記事から，各漢字の出現頻度を調査したが，より一般的で説得力のある重要度決定方法を考案する必要がある。現在考えている一手法は，図 2や図 3のシミュレーションモデルを作り，データの送信にかかる時間などを仮定した上で，解析を行なう。そしてクライアントの待ち時間が最小となるようなグループ分けの方法を導き出すという手法である。

フォント情報交換の国際規格として，ISO/IEC 9541 がある。この規格では，フォント情報内容そのものを規定するのではなく，フォント情報の構造と表記方法を規定して，フォント情報交換の必要条件の充足を図っている。インターネット上で利用できるフォントサービスを構築する場合，この規格に適合したもので，なおかつマルチプラットフォームで実現しなければならないと考えられる。

次に，フォントが必要となるたびにネットワークを介してサーバから取ってくるという方法では，効率が悪い。そこで，フォント・キャッシュ・システムが必要となる。ここでは，通常のキャッシュ・システムとは異なり，フォント情報は頻繁に更新されたりすることがないという点を考慮しなければならない。また，それにともなって，フォントサーバやキャッシングプロキシの最適配置問題も生じる。

また，フォントは知的所有権を伴う情報であるので，実現に当たっては，ライセンス問題や課金システムなどの検討が必要となる。

参考文献

[1] 小町裕史 : 文書記述言語の標準化動向 – V フォント情報交換の国際標準化，情報処理，**34** No. 7, pp. 915–921 (1993).