WWWにおける先読みページ選択手法に関する一提案

5F－3

金子郁夫　松永賢次
専修大学　経営学部

1.研究の目的

Web 利用時の待ち時間を軽減することを目的とする。ユーザーがネットを利用しページをブラウズしている間，回線にはデータが流れていない。この無駄を有効に利用することによって待ち時間の軽減が図れると思われる。その方法の一つとして，ユーザーのアクセスを予測し，先読みをするシステムを作成する。

このシステムの鍵を握るものは先読みのための予測情報である。この予測の精度で先読みが有効であるかが決定されるため，いくつかの方法で予測情報を作り出し，精度の評価をする。

2.システム全体のアーキテクチャ

このシステムはプロキシーに常駐するものである。(図 1)

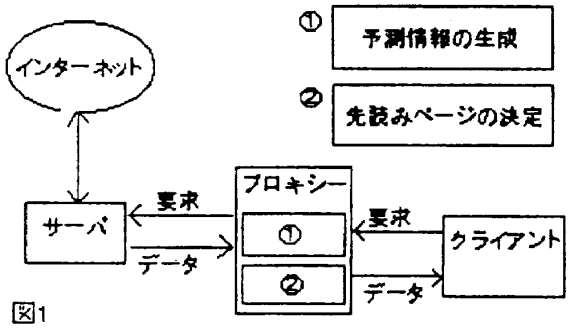

図1

クライアントから要求を受けると，プロキシーでは要求されたページに関する依存関係表を参照し，優先順位とその他の情報を考慮し，最終的な選択をして先読みを開始する。

クライアント側では，Netscape などのブラウザを使用し，このシステムのために設定を変える必要は無く，普段通りに使用する。

3.予測情報生成のアルゴリズム

3.1 予測情報生成の情報源

取得したファイルのリンクをたどる方法とアクセスログを使う方法に大別できる。本研究では予測情報の生成にはプロキシー内に残されるアクセスログを使用することにした。アクセスログから，ページ間の依存関係を調べ，あるページに対する別のページの出現頻度を計算し先読みの優先順位を決定する。

アクセスログをソースとする先読み対象の決定におけるメリットは，事前に依存関係表が出来上がっていることで，ファイルの大きさや，先読み対象ページまでの距離などから読み込みにかかる時間を事前に調査しておくことができるところである。

一方，デメリットとしては，依存関係の表にどれだけ有効な先読み対象が含まれているかという問題があげられる。アクセスログ内のページの出現順で依存関係を作ると，機械的な処理のため無駄な依存関係がたくさん作られてしまう可能性が大きく，これらのゴミをどのように判断し処理するかを決定するのはアクセスログだけからは非常に難しい。

3.2 依存関係の範囲と強さ

3.2.1 表の作成

依存関係の表を作るためにはどこまで後ろの読み込みファイルまで関係があるか決定しなければならない。また関係があるとみなしたファイルの重み付けも決定しなければならない。そ

A proposal for prefetching method on WWW
Ikuo Kaneko and Kenji Matsunaga
Senshu University
2-1-1 Higashi-mita, Tama-ku, Kawasaki, Kanagawa
214-8580, Japan

のためにもっとも単純な方法である，順番のみを考慮して依存関係を作り出すという方法に基づいてプログラムを作成し，専修大学のプロキシーサーバが出力したアクセスログを解析し調査してみた。

この際使用したログからは，画像ファイル，cgi ファイル，特殊な記号を含んだログ，などを事前に削除して，なるべく html ファイルへのアクセスのみを残すようにした。アクセス数は約 1 万件である。

親ページからどのくらいまでの範囲が依存関係をもっているかを調べるために，5，10，20 ページの範囲で得点を付けた。

3.2.2 調査の結果

得点の最大点は範囲に決めたページ数と同じにし，得点差は 1 点とした。このとき，範囲を 5 ページに設定した依存関係表は 4 万件弱になり，10 ページでは約 7 万，20 ページでは約 13 万件になった。

範囲	5	10	20
依存関係表の件数	4万	7万	13万

Yahoo!Japan のメインページを例にとってみると，微妙に差が出ていることが分かる。数的には差が無いように見えるものの，同サイトへの優先順位が高くなっているのはどの場合であっても，範囲の最も小さい 5 ページと設定しているときである。逆に範囲を大きくすると関係が薄いと見て取れるログが上位に浮かんできたりすることもあり，この単純に順位で決定する場合においては，上限は 10 ページ位が適当ではないかと思えた。

このことから，時間を考慮して依存関係を作成する上で適した形だと推測されるものは，親ファイルからある程度の間までは得点差が無く，多少距離が空いてくると得点の減少の割合が大きくなるという形が，親ファイルからの依存関係を強く表現することができると現段階では考察される。

4 考察

4.1 他の手法との併用

ログの中で同じページが複数回出現する確率は 20%ぐらいであった。ページのリンクから先読みページを選択する方法では余分な労働を 80%分発生させていることになる。ページのリンクから先読みページを選択する方法とアクセスログを使用する方法と併用すればユーザーは確実にリンクをたどれる。

ファイルの大きさや更新日時を事前に調べておき，ネットワークの状態と負荷を考慮しつつアクセスログを使う方法は事前に表が作られているので先読み対象を決定することも可能になる。

4.2 依存関係表の大きさについて

5，10，20 ページの範囲で依存関係を作ったものはそれぞれ，6，11，21 点以上の点数が付けられている子ファイルが親ファイルと関係していると考えられ，それぞれ 3500 件，7300 件，15400 件が親ファイルと関係があると見て取れる。アクセスされなくなったページやアクセスの少ないものを削除することで，データ量が多くなりすぎないようにすることが可能と考えられる。

参考文献

1) Padmanabhan, V. and Mogul, J. : Using Predictive Prefetching to Improve World Wide Web Latency, ACM SIGCOMM, Vol.26, No.3, pp22-36(1996).
2) 知念賢一，岡山聖彦，山口英：WWW におけるインタラクティブな先読みシステムの設計と実装，コンピュータソフトウェア，Vol15,No2,pp48-61(1998).