マルチメディア通信と分散処理 77－10
（１９９６．７．１２）

WWW 先読み代理サーバにおける先読み対象決定戦略

知念 賢一, 山口 英
奈良先端科学技術大学院大学

概要

Internet の普及にともない高速な情報サービスが求められており、その高速化の手段として先読みを実現したシステムが研究・開発されている。情報サービスの先読みを実現するシステムの性能は、先読み対象決定戦略に大きく影響される。本研究では WWW において先読みを実現する先読み代理サーバを実装した。本論文では、WWW 先読みサーバにおける先読み対象決定戦略について述べる。そして、実在する情報と要求の解析による、先読み対象決定戦略の最適法について述べる。

キーワード：WWW、先読み、先読み対象決定戦略

A Prefetching Strategy on Prefetching Proxy Server for WWW

Kenichi Chinen and *Suguru Yamaguchi*
Nara Institute of Science and Technology

Abstract

As the Internet becames popular, needs for high speed information services have been growing. Many prefetching systems are designed and implemented for high speed information services. Performance of prefetching system is affected by its strategy. This paper describe prefetching strategy on prefetching proxy server for WWW and optimization by analysis of actual information and requests, and report on the implementation.

Keywords: *WWW, prefetching, prefeching strategy*

1 はじめに

近年Internetが普及し、新たな情報基盤として注目されている。Internetは様々な特性をもつネットワークの集合体であることから、サーバやクライアントが接続されたネットワーク、および中継するネットワークの影響をうけて通信時間の変動や遅延が生ずる。Internet上の情報サービスを高速に提供するには、通信時間の変動や遅延を回避する技術が必要である。

従来より、通信時間の短縮や情報サービスの高速化は、高速・広帯域なネットワークの導入によって実現されることが多かった。同様の手段でInternetのサービスを高速化するには、Internetへ高速・広帯域なネットワークの導入が必要である。しかし、Internetは多種多様のネットワークが様々なポリシーで接続されているため、サーバやクライアントの接続されるネットワーク、とりうる経路の全てのネットワークを高速・広帯域化することは非常に困難である。したがって、Internet上のサービスの高速化は高速・広帯域なネットワークの導入とは別の手段で高速化を実現せねばならない。

本研究では利用者の要求を待たず、あらかじめ情報を転送する先読み技術(prefetching technique)に注目し、Internet上の代表的な情報サービスのWWW (World-Wide Web)を高速化するための先読み代理サーバを設計・実装した[1]。先読み代理サーバを評価することにより、先読みがWWWサービスの高速化に効果があることが明らかにした。また、先読み代理サーバの実装・評価を通して、先読みを実現するシステムの性能には、先読み対象を決定する戦略が大きな影響を与えるとの知見を得た。

本論文では先読み代理サーバにおける先読み戦略を検討する。また、実在する情報と要求を解析し、先読み対象決定戦略を最適化する。

2 先読みと先読み対象決定戦略

Internet上の情報サービスにおける先読みは、利用者の要求を待たずにあらかじめ情報を転送する技術である。先読みを実現するシステムでは、利用者の要求を自動的に推測して先読みを行い、利用者の情報へのアクセス時間を短縮することを目的とする。一方、先読みを実現したシステムの性能を考えた場合、利用者が利用しない情報を転送しては資源の浪費となり、利用者が利用する情報を転送しない場合には先読み効果が得られない。したがって、利用者がどの情報を要求するかを的確に推測することが重要である。また、時間と資源は有限であるから、要求されると推測した先読み対象が多い場合には、システムの環境に応じて先読み対象を選択せねばならない。

本論文では先読み対象の推測方法と選択方法をあわせて「先読み対象決定戦略」と呼ぶ。

2.1 先読み対象の推測

利用される情報の推測は連想的推測と統計的推測に大別される。前者は利用者の要求した情報の内容から次の要求を推測する方法で、情報の内容から次の要求を推測できる場合に用いる。特にNetNewsのヘッダやHTMLの参照等、利用者の要求した情報に関連する情報についてのヒントが含まれている場合に有効である。後者は統計的手法により利用者の要求する傾向を調査して、次の要求を推測する方法である。アクセス頻度等の統計的手法によって利用者の要求する傾向が把握できる場合に有効である。また、これら2つの手法を組み合わせることも考えられる。

2.2 先読み対象の選択

先読みシステムの稼働するコンピュータ環境の資源は有限であるため、推測結果の全てを先読みして保持することは困難である。また、短時間に先読みを行えば大きな先読みの効果が得られるが、先読みは大量のプロセスやトラフィックの発生等、大きな負荷が予想

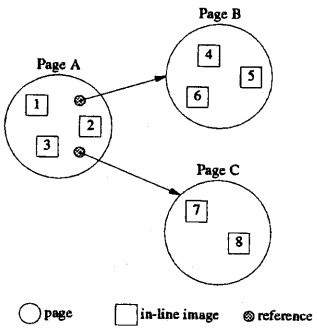

図 1: ページでみる情報の提供形態

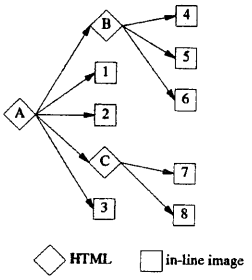

図 2: 個々の情報でみた情報の提供形態

され、短時間に多くの情報を先読みすることは困難である。以上の理由から、要求されると推測した情報の中から、先読み対象とする情報を選択する必要がある。

3 WWW における先読み対象決定戦略

これまで、WWW における先読みはほとんど研究されず、先読みを実現したシステムの挙動、どのような先読み対象決定戦略が可能なのか明らかにされていなかった。本章では本研究で設計・開発した WWW における先読み対象決定戦略について述べる。

3.1 WWW における推測

WWW のサービスにおいては、利用者がクライアントを対話的に操作して、情報を得ることが多い。したがって、WWW の先読みはバッチ処理ではなく、利用者の利用に応じて逐次的に全ての情報の転送を先読みすることが望ましい。前述のように、先読みには連想的推測と統計的推測による先読みが存在するが、統計的推測は履歴の記録など、事前に必要とされる処理が多く、有意な推測効果を得るのには比較的時間を要する。一方、連想的先読みは統計的推測に比べ必要とする処理が少なく、比較的短時間に推測できる。また、後述するように、WWW においては HTML (HyperText Markup Language) [2] の記述を走査によって連想的な推測が可能である。これらの点から、本研究では連想的推測を基に先読み対象決定戦略を設計した。

WWW のサービスはページと呼ばれる単位で提供されている (図 1)。そして、ページは一つの情報、あるいは HTML で記述されたハイパーテキストである。ハイパーテキストには、ページに含まれる情報や他のページへの参照を URL (Uniform Resource Locator) [3] によって記述する (図 2)。HTML を走査して URL を取り出すことによって、参照ページやページに含まれる情報の推測が可能である。利用者によるアクセスの多くは URL で指定された参照に沿って他のページを要求することであり、参照ページを先読みすることがアクセス時間の短縮となる。参照ページのアクセス時間を短縮するには HTML によるページの記述だけでなく、参照ページに含まれる情報も先読みする必要がある。参照ページに含まれる情報の先読みは、特にインラインイメージと呼ばれる画像情報がページに含まれている際に有効である。

3.2 WWW における選択

前述のように各利用者が利用できる時間や資源 (帯域、処理能力) は有限であるため、推測された情報が大量に存在する場合には先読み対象を選択する必要がある。どのような情報を選択することが効果的な先読みかを検討せねばらならない。

まず、参照される情報の属性から選択する方法を検討する。アクセス時間を短くする点を最優先に考えた場合には情報のサイズが、新たな情報を得る点を最優先に考えた場合には、情報の更新時刻あるいは破棄時刻が対象を選択する重要な要素である。また、画像やテキストのような情報の型を知ることができれば、型に応じた選択方法が設計可能である。しかし、WWW で情報を URL からその URL が示す情報のサイズ、更新時刻、破棄時刻を取得することは不可能である。また、一部の型の URL の末尾の拡張子による類推を除いて、型の取得も困難である。したがって、本研究では情報の属性による対象の選択は実現できなかった。なお、情報のサイズ、更新時刻、型はサーバに問い合わせることによって取得が可能だが、大量の先読み対象を選択する時点で、全ての情報の特徴をサーバに問い合わせるのは多くの時間を必要とし、本研究の目的である高速化には適さない。

本研究では WWW のクライアントの挙動から先読み対象の決定方法を検討した。多くのクライアントは HTML の記述を評価し、HTML に記述されている順に情報を要求する。それらの情報のうち、ページの先頭にはページの説明やページを印象づける画像等が多く、提供者・利用者ともにもっとも高速に得たい情報であると考えられる。つまり、ページに含まれる情報は HTML に記述されている位置に応じて重要さが異る。このことから、推測結果が大量な場合、先頭から数個を先読み対象として決定する。先頭から何個を先読みするかは、実際のシステムの実装時に定めるものとする。

4 先読み代理サーバによる実装

クライアント、代理サーバ以外に新たなシステムを導入することは困難であること、前述のように先読みの実現は大きな負荷やトラフィックが予想されことから、本研究では代理サーバにおいて先読みを実現した。先読みサーバで先読みを実現することにより、先読み結果の共有と先読みによる LAN 内のトラフィックの増加を抑制した。先読みの機能を付加した代理サーバ(以降、先読み代理サーバと呼ぶ) では、クライアントへ HTML で記述した情報を転送する際に先読みを行う。ここでは、この HTML による記述で構成されるページを「基ページ (base page)」と呼ぶ。先読み代理サーバは、1) 参照ページの HTML による記述の先読み、2) 参照ページに含まれる情報の先読み、の 2 段階で参照ページの先読みを行う。参照ページの推測は基ページの HTML による記述の走査により、参照ページに含まれる情報の推測は参照ページの HTML による記述の走査による。

5 要求の解析と選択方法の最適化

本研究で設計した先読み対象決定戦略の性能は、先読み個数をパラメータとして変動する。本章では、実際に運用される先読み代理サーバで、効率のよい先読み対象決定戦略のパラメータを導く手法として、利用者の要求を解析する。解析に用いた利用者の要求は、奈良先端科学技術大学院大学のキャンパスネットワークで運用されている代理サーバのログから 1 日分 (8519 個) を取り出したものである。

5.1 推測される情報数

前述のように、WWW では HTML によるページの記述を走査することによって、先読み対象の推測が可能である。要求されたページを実際に走査して参照ページを取り出したところ、参照するページは最小 0 個、最大 1272 個であった。そして、要求されたページと参照するページを走査して、ページに含まれる情報を走査したところ、含まれる情報は最小 0 個、最大 424 個であり、基ページから推測した参照ページに含まれる情報の数は最小 0 個、最大 2011 個であった。先読みシステムでは利用者の要求が発生から次の要求を推測して、次の要求の到着までの短時間に情報を通

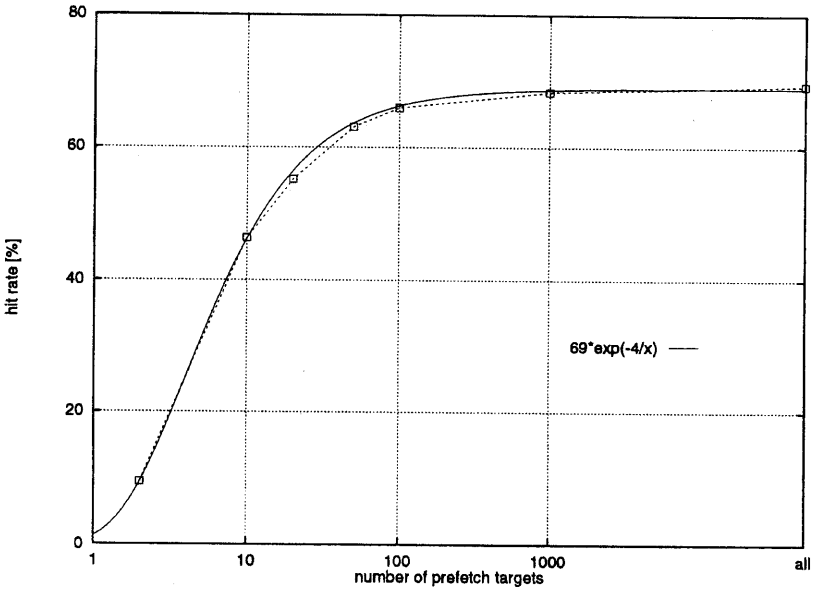

図 3: 先読み個数とヒット率の相関

信する必要があるが、短時間に数十から数千の情報を通信するのは困難であり、この状況では先読み対象が多い場合には選択が必要である。

5.2 先読み個数の算出

効果的な先読み個数を求めるため、先読み個数をいくつか設定して先読みのシュミレーションを行なった。シュミレーションは実装が容易な代理サーバにおけるヒット率の算出を行った。ヒット率が高い場合にはクライアントへのレスポンスが向上することから、高いヒット率は短いアクセス時間で情報を転送したものとみなせる。この場合、クライアントから要求された情報を基ページとして、参照ページと参照ページに含まれる情報を推測し、それらの情報がクライアントから要求された場合をヒットとする。ヒットする回数とクライアントからの要求の回数の比がヒット率である。先読み代理サーバではキャッシングの拡張として先読みを実装しているが、このシュミレーションではキャッシングが先読みの効果へ影響しないよう、過去の要求の履歴を用いず、推測結果の履歴のみを用いた。よって、先読みの効果によるヒット率のみ算出される。

その結果、先読み個数が大きくなるに従いヒット率が向上することが明らかとなった(図3)。特に、推測される先読み対象を全て先読みした場合には、68.4% のヒット率が得られた。また、先読み個数とヒット率は線形ではなく、膨大な情報を先読みしても容易にヒット率が向上しないことが明らかになった。

シュミレーションで得られらた結果の近似式として、次式が得られた。

$$y = 69 * e^{-4/x}$$

この式より最大性能の 90%、99% を出すために必要な先読み個数を求めると、以下のようになる。

$$x = \frac{4}{\ln(69) - \ln(y)}$$

$$\frac{4}{\ln(69) - \ln(69 * 0.90)} = 37.965$$

$$\frac{4}{\ln(69) - \ln(69 * 0.99)} = 397.997$$

すなわち、最大性能の90%、99%を出すためには、それぞれ38個、398個先読みする必要がある。よって、38個以上を設定すれば、大きな性能をもった先読み代理サーバとなり、398個以上を設定すれば、ほぼ最大の性能をもった先読み代理サーバとなる。

6 今後の課題

本研究ではシュミレーションにより、連想的推測に基づく先読み対象決定戦略を用いる際の最適な先読み数を導出した。しかし、導出された先読み数は数十から数百であり、単一のシステムで処理するのは困難である。したがって、数十から数百を処理できるシステムの実装が課題である。また、Harvestのような複数のシステムが協調する代理サーバへの実装の検討も必要であろう[4]。

本研究では利用者の要求の解析とシュミレーションで先読み対象決定戦略の性能を検討したが、実際のシステムの性能は利用者の別の挙動も考慮する必要がある。例えば、多くのクライアントのもつreloadと呼ばれる機能は、代理サーバへサーバから新しい情報の転送を促す機能であり、ヒット率の低下をもたらす。このような、実際にシステムで起こる現象を考慮した評価はシュミレーションでは困難である。したがって、実際のシステムの稼働による評価が必要である。

現実のWWWサービスの利用では、特定の情報に利用が集中していることが広く知られている。したがって、今後の先読みシステムの研究は、連想的推測だけでなく統計的推測に基づいた先読み対象決定戦略も必要である。そして、統計的推測に基づく先読み対象決定戦略の視点で、要求を解析して検討する必要がある。

7 おわりに

情報サービスの高速化手段として先読みを用いた場合、そのサービスを提供するシステムの性能には先読み戦略が大きく影響する。WWWの利用形態に応じた先読みを行うためには、代理サーバにおいてクライアントの要求するHTMLの記述を走査して先読み対象を決定する連想的先読みが効果的である。

実際のクライアントからの要求に基づいてInternet上で提供されている情報を解析することによって、WWWの利用者の要求の多くは連想的推測で推測可能であることを示した。また、シュミレーションによって先読み個数に応じてヒット率が変動すること、最適な先読み個数を求める手法を明らかにした。

謝辞

多くの支援をいただいた奈良先端科学技術大学院大学 情報ネットワーク講座の皆様に感謝いたします。特に同講座の岡山助手には多くの協力を承りました。心から感謝致します。

参考文献

[1] 知念賢一, 山口英. 先読みによるWWWアクセスの高速化の可能性. インターネットコンファレンス'96, July 1996.

[2] T. Berners-Lee and D. Connolly. Hypertext Markup Language - 2.0, RFC1866. November 1995.

[3] T. Berners-Lee. Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web, RFC1630. June 1994.

[4] Anawat Chankhunthod, Peter B. Danzig, Chuck Neerdaels, Michael F. Schwartz, and Kurt J. Worrell. A Hierarchical Internet Object Cache. *USENIX 1996 TECHNICAL CONFERENCE*, January 1996. <URL:http://excalibur.usc.edu/cache-html/cache.html>.