社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CG－125 (12)
2006／11／16

可視化による大容量 Web プロキシサーバログ解析システム

坂本良平† 瀬川大勝†‡ 宮村（中村）浩子† 斎藤隆文†

†東京農工大学 大学院生物システム応用科学府

‡東京農工大学 総合情報メディアセンター

本稿では大容量 Web プロキシサーバログの解析を支援するための可視化手法を提案する．Web プロキシサーバのログを解析することによって，アクセス量，キャッシュ状況などの実際の動作を把握することができる．しかし，従来のログ解析ツールの表示方法では，画面の制限などの理由によって表示可能データが限られるため，ログデータの集計結果を表示するにとどまり，詳細な動作の解析には限界がある．本稿では大量のログデータをできる限り総覧表示することで，Web プロキシサーバの詳細な動作を効率的に把握できるような可視化手法を提案する．また，本システムでは膨大なデータの中から有益な情報を得るビジュアルデータマイニングを考慮し，ユーザが可視化結果に対して対話的にソートやドリルダウンができる機能を提供する．

Web Proxy Server Analysis System Using a Visualization of the Huge Log

Ryohei SAKAMOTO†, Hirokatsu SEGAWA†‡

Hiroko (NAKAMURA)MIYAMURA†, and Takafumi SAITO†

†Graduate School of Bio-Applications and Systems Engineering,
Tokyo University of Agriculture and Technology

‡General Information Media Center, Tokyo University of Agriculture and Technology.

In this paper, we propose an analysis system on a Web proxy server by visualizing huge log data. By analyzing log data, we can comprehend the actual status of traffic, cache, and so on. By using the existing tools, however, can only show the summary at a same time because the log data size is too large to be displayed on the computer screen. That is why it is difficult to analyze the detailed status. Our visualization technique enables to display the lots of data on the screen and to get the detailed information of the Web proxy server effectively. For visual data mining, there are several interactive operations we also provide.

1. はじめに

本稿では大容量 Web プロキシサーバログの解析を支援するための可視化手法を提案する．

LAN 内のホストからインターネットの Web サイトにアクセスする際，Web プロキシサーバを経由することにより，キャッシング，フィルタリング，匿名性確保などの効果が期待できる．これらの Web プロキシサーバの特性から，ある程度規模の大きいネットワークでは，高速化やセキュリティの向上を目的として Web プロキシサーバを導入するケースが多い．

ネットワークの管理者はその効果的な性能を保つため，Web プロキシサーバに記録されるアクセスログを解析する必要がある．また，アクセスログからネットワーク上の新たな問題を発見できる可能性がある．しかし，企業や大学などの大規模なネットワークではログの量が膨大となり，テキストで記述されたログを解

析するのは困難かつ多大な手間を要する．そのため，ログから有益な情報をわかりやすく提示するツールの需要は高く，ログの解析ツールは多く開発されており，フリーウェアや企業による解析サービスも多い．しかし，これらのサービスによるレポートはアクセスログを集計しただけのものが多く，結果表示は単純なグラフやテキストの羅列である．その結果，レポートから有益な情報を得るのにはやはり手間がかかってしまう場合がある．特に，多くのデータを比較したい場合などは一画面に多くの情報を表示する必要があるが，画面の領域制限のため限界があるため，解析は困難なものとなる．

そこで，本稿ではWebプロキシサーバのアクセスログを可視化することによって，効果的な解析を支援することを目的とした手法を提案する．

2. 関連研究

解析者が得たいと考える情報によって使用するデータが異なり，可視化手法も異なる．これまでにログ解析の可視化手法として，ログのURLを元に取得したHTMLファイルからキーワードを抽出し，Webユーザのアクセス動向を把握するための可視化[1]やログメッセージの出力パターンに着目して問題の兆候を発見するための可視化[2]が提案されている．本稿では，ログのキャッシュオブジェクトの属性に着目し，キャッシュ機能を最適化するための解析を支援することを目的としている．

3. 提案手法

本章では提案手法について述べる．入力データとなるWebプロキシサーバのアクセスログについて，および解析の対象であるキャッシュ機能の最適化について述べた後，提案する可視化手法について述べる．

3.1 アクセスログの取得

一般的にWebプロキシサーバのアクセスログはテキストで記されている．図1は代表的なプロキシサーバ『squid』のアクセスログの内容である．アクセスログにはユーザがWebサイトにアクセスを要求した時刻，オブジェクトの取得時間，送信元ホストのIPアドレス，ログタグ，オブジェクトのサイズ，アクセス先のURL，ファイル形式などが記されており，Webユーザが Webページにアクセスを要求する度にログが追加される．プロキシサーバは一定期間のアクセスログを一つのファイルとして出力することが一般的であるが，大規模なネットワークにおけるログファイルは一日あたり数百万行を越える膨大なテキストになる場合もある．このログファイルを入力データとして可視化を実現する．

```
1125760014.117  610  192.168.11.82  TCP_MISS/200  45338
 アクセス時間  取得時間  IPアドレス     ログタグ       サイズ

GET http://www.yahoo.co.jp/ - DIRECT/ *.*.8.46  text/html
メソッド      送信先URL                          ファイル形式
```

図1: Webプロキシサーバ(squid)のログ

3.2 キャッシュ機能の最適化

本研究で使用するプロキシサーバであるsquidはキャッシュ機能を備えている．インターネットのWebサーバから一度取得したオブジェクトをプロキシサーバに保存しておくことにより，再び閲覧する際はインターネットにアクセスすることなくWebプロキシサーバからオブジェクトを取得することができる．その結果，取得時間は短縮され，トラフィックも軽減されることが期待できる．しかし，キャッシュ容量には制限があるため，実際はLRUアルゴリズムなどによってオブジェクトの入れ替えがおこなわれている．ディスクにオブジェクトが存在しない（ヒットしない）場合，逆にWebプロキシサーバを経由した分遅延が生じてしまう．よって，ヒット率を高く保つことが重要な課題である．

プロキシサーバは基本的に取得したオブジェクトをすべてキャッシュするが，設定によって指定したオブジェクトをキャッシュしないようにすること可能である．キャッシュしないことが望ましいとされるオブジェクトは次のようなものである．

(A) 取得時間が大きいオブジェクト
(B) 更新頻度が大きいオブジェクト

(A) 取得時間が大きいオブジェクト

オブジェクトによって，Webサーバから直接取得する場合とキャッシュを利用する場合で取得時間の差が小さなものと大きなものがある[3]．キャッシュによる効果を最大限にするためには，後者のようなオブジェクトをできる限りキャッシュに残しておくようにし，前者のようなオブジェクトをキャッシュしないように設定することが望ましい．

(B) 更新頻度が大きいオブジェクト

squid はオブジェクトに年齢の概念を持たせている．これはユーザに古いオブジェクトを提供しないようにするための機能であり，オブジェクトがヒットした場合でも，そのオブジェクトが古ければ Web サイトから新鮮なオブジェクトを取得する．Web サーバによって動的に作成されるオブジェクトや更新の早いオブジェクトはすぐに新鮮ではなくなるため，キャッシュしても効率がよくない．従って，上記のような更新の早いオブジェクトもキャッシュしないようにすることが望ましい．

上記のようなオブジェクトをキャッシュしない設定をおこなうためには，取得時間や更新頻度の早さなど，オブジェクトの属性を知る必要がある．また，上記のような望ましくない特徴があっても，アクセス頻度が微々たるものであれば問題とならないため，アクセス頻度も重要となる．そこで，可視化をおこなう場合，オブジェクトのアクセス頻度，取得時間，更新頻度の早さなどの属性値を複合的に判断できるレイアウトを考える．

3.3 可視化手法

本節では可視化手法について述べる．基本的なアイデアであるオブジェクトのリスト表示について述べた後，効率的に閲覧するための短縮法・ドメイン表示法，対話的操作について説明する．

3.3.1 オブジェクトのリスト表示

3.2項で述べたように，オブジェクトのアクセス数，取得時間，更新の早さなどのさまざまな属性を複合的に見てキャッシュの設定を決定したい．そのための可視化手法を実現する．図2に本手法の概略図を示す．

画面にはアクセスログから集計した URL，アクセス数，サイズ，取得時間，ヒット数，ファイルの種類，取得メソッドなどの属性を並列に配置したオブジェクトのリストを表示する．これにより，一画面に表示されているオブジェクトの各属性を一度に見比べることが可能である．また，アクセス数や取得時間などは数値データであり，表示データ数が多い場合，または表示サイズが小さい場合は読み取るのが難しい．そこで，数値の大きさによって色をつけることで直感的になり，おおまかな数値の読み取りが容易となる．

オブジェクトの取得時間やサイズなどの数値データはアクセスログから読み取ることができるが，更新の早さは数値で記載されていない．しかし，ログタグに着目することで更新が早いオブジェクトを推定することができる．ログタグは，Web プロキシサーバがどのような処理をしたかを示す情報である．表1に主なログタグの種類を示す．squid はオブジェクトに年齢の概念を持たせることによって，ユーザに古い情報を提供しないようにしている．ログタグの『TCP_REFRESH_HIT』はキャッシュにオブジェクトは存在していたが，新鮮なデータではなかったため新しいデータを Web サイトから取得したことを表している．つまり，『TCP_REFRESH_HIT』が頻繁に出現するオブジェクトはデータの更新頻度が大きいことが推測できる．

『TCP_IMS_HIT』はクライアントのコンピュータが所持するキャッシュにそのオブジェクトを持っている場合の要求で，クライアントのキャッシュオブジェクトよりも新しいものが Web プロキシサーバに存在した場合，そのオブジェクトをクライアントに渡します．このように，ログタグからオブジェクトの新鮮さの情報を得ることができる．そこで，オブジェクトの属性値と同時にログタグの情報を表示する描画領域を設ける．この描画領域にはオブジェクトにアクセスがあった時間とそのアクセスに対するログタグを表示する．

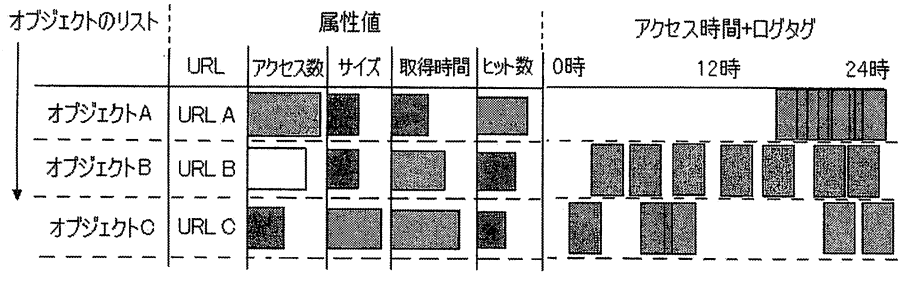


図2：可視化手法（オブジェクトの属性値とログタグの総覧）

表 1：主なログタグの種類

ログタグ	処理	色
TCP_HIT	オブジェクトが存在した	赤
TCP_MEM_HIT	オブジェクトがメモリに存在した	橙
TCP_REFRESH_HIT	存在したが新鮮ではなかった	緑
TCP_IMS_HIT	オブジェクトが存在した (IMS 要求)	青
TCP_MISS	オブジェクトが存在しなかった	灰

横軸に時間軸をとり，一アクセス毎に一つの長方形を描画する．そして，その長方形をログタグの種類によって色付けする．この表示方法により，各オブジェクトへのアクセスがどれくらいの頻度でおこり，どのような処理がされているかの概要を容易に把握することができる．

3. 3. 2 短縮表示・ドメイン別表示

大規模なネットワークにおいて Web サイトへのアクセス量は膨大であり，Web プロキシサーバには大量のアクセスログが残る．ログに記録されている全てのオブジェクトを前項で述べたリストとして表示することは現実的ではない．多くの場合，注目すべき箇所はアクセスの多いオブジェクトであるため，アクセス数の多いオブジェクトのみを表示すればよい．しかし，それでも多くのオブジェクトが表示され，データの閲覧に時間がかかってしまう．さらに効率的よく閲覧するため，オブジェクトの特徴を利用して，情報量を減らすことなくできる限り表示データ量を削減する．Web ページを構成しているオブジェクトは Web サーバのファイルシステムにおいて階層的に管理されている．同じディレクトリに存在するオブジェクトは取得時間，サイズ，アクセス頻度などにおいて同じような特徴を持つことが多い．そこで，同じ階層にある（URL のファイル名だけが異なる）オブジェクトがリスト上で連続する場合，それらの中の一つのみを表示する機能をつける．

オブジェクトをキャッシュしないという設定をおこなう場合，文字列を指定する．その文字列を含んだオブジェクトはキャッシュされなくなるが，多くの場合は文字列としてドメインや URL を指定する．キャッシュしないことが望ましいオブジェクトを多く含むドメインを見つけることができれば効率的である．そこで，ドメインごとにオブジェクトを表示する機能をつける．

3. 3. 3 対話的操作

対話的操作によって有益な情報まで効率的にアクセスできる場合がある．本稿ではソートとドリルダウンを実装する．各属性の値によってリストをソートすることで，属性値が大きいオブジェクトを見つけるだけでなく，属性間の相関を発見することができるなど，有益な情報が得られる可能性がある．また，オブジェクトのリスト表示から着目したオブジェクトを選択操作によりドメイン表示に切り替えることで効率的に得たい情報にアクセスできる．

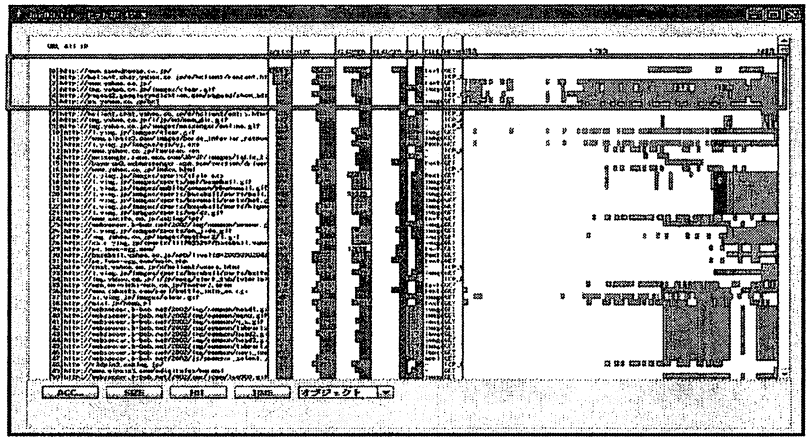
4. 可視化結果

本節では，実際のアクセスログを適用した可視化結果を示す．適用するアクセスログはある大規模ネットワークに設置されている Web プロキシサーバ(squid)のものであり，ログファイルの容量は約 50MB，テキストの行数にして約 50 万行である．

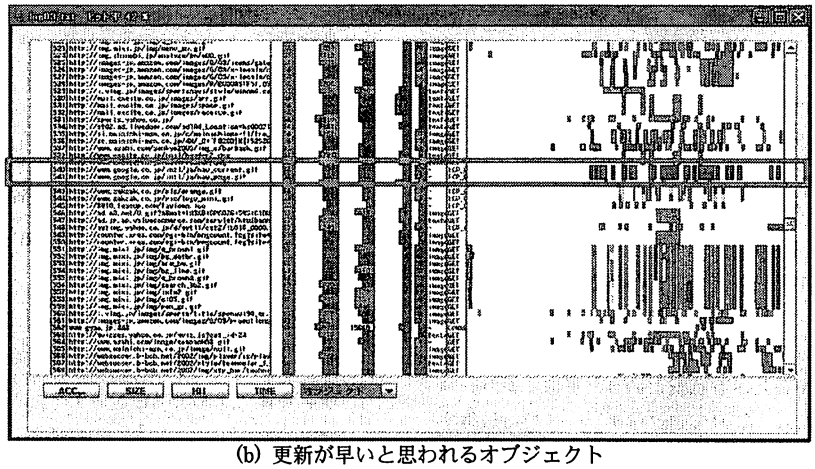
可視化結果を図 3 に示す．属性値は値の小さい方から青→緑→黄→赤で色が割り当てられている．また，ログタグの種類を表す長方形の色の分け方は表 1 のとおりである．

図 3(a)はアクセス数の大きい順に降順でオブジェクトのリストが表示されている．各属性値を示す領域(画面中央辺り）からは，取得時間やサイズの大きいものを容易に理解することができる．注目すべきはアクセス数が多いオブジェクトがほとんどヒットしていないことである．アクセス頻度が大きいオブジェクトほどキャッシュを利用した方が効果的であるため，原因を解明してヒットさせるようにすべきである．

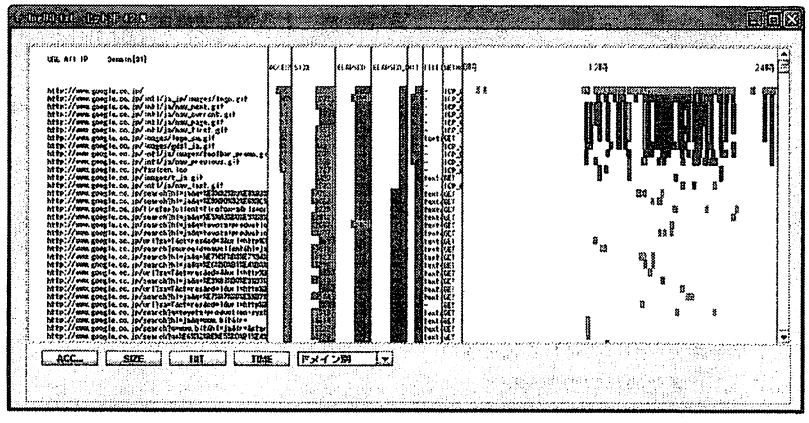
図 3(b)の赤枠で囲まれているオブジェクトは，アクセスが連続でおこなわれているがログタグはすべて TCP_REFERESH_HIT であることがわかる．従って，このオブジェクトは更新が早いものであると推測される．また，そうした特徴に着目してオブジェクトのドメイン表示に切り替えると，同ドメインのオブジェクトはほとんど同じ特徴を持っていることがわかった（図 3(c)）．その結果，このドメインをキャッシュしない設定の文字列として指定することによってキャッシュ資源をより効率よく使用できる可能性がある．



(a) アクセスは多いがヒットしていないオブジェクト

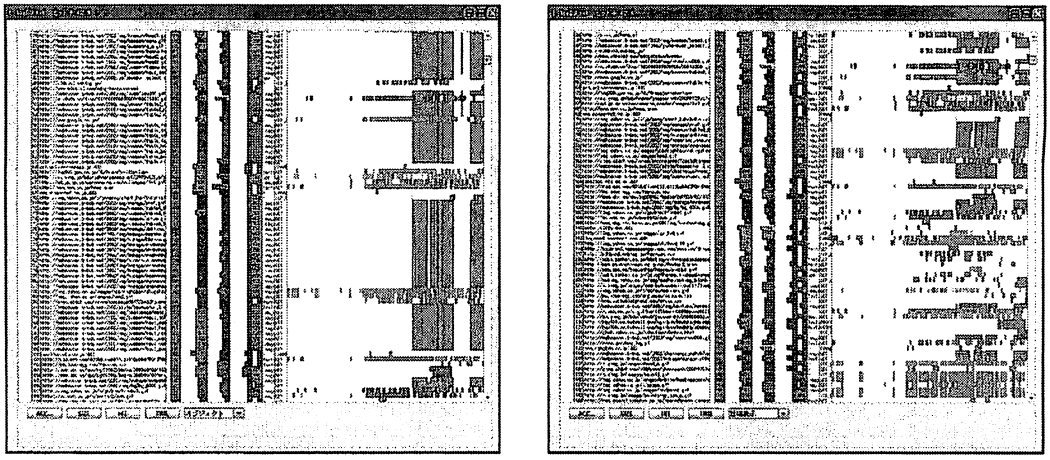


(b) 更新が早いと思われるオブジェクト



(c) 更新が早いと思われるオブジェクトのドメイン表示

図 3：結果画像



(a) 短縮前　　(b) 短縮後

図 4：短縮表示による表示データ数の削減

短縮表示を適用したことで，30%程度データ表示数を削減でき，効率的にリストを閲覧できるようになった（図 4）．また，各属性値でリストをソートすることによって，サイズが大きいオブジェクトは取得時間も大きいといった属性間の相関や，サイズが大きいものは動画ファイルが多いなどといった特徴を理解することができる．

その他，一日中定期的にアクセスされているオブジェクトの発見や各ヒット状況といったプロキシサーバの詳細な動作など，従来のログ解析ツールでは取得するのに手間のかかっていたさまざまな情報を可視化することで容易に得ることができた．

5. おわりに

本稿では，大容量 Web プロキシサーバのログ解析を支援するための可視化手法を提案した．本手法はオブジェクトの属性値，ログタグなどを同時に一画面に表示することで複合的な解析が可能となった．その結果，管理者がキャッシュの設定を改善するための判断材料となりうる．

今後の課題としては，今回使用していないネットワークユーザなどのアクセスログから得られる情報を追加することによって，より多角的な解析を実現し，キャッシュ設定において適切な判断を可能にしたい．また，有用なデータをさらに効率的に取得するためのデータ表示法や対話的操作を考える．

参考文献

[1] 戸川聡，金西計英，矢野米雄，Web 閲覧特性に基づく管理者支援のための利用動向可視化システム，情報処理学会論文誌，Vol. 46, No. 4, pp. 985-994, 2005.

[2] 高田哲司，小池英樹，見えログ:人間による計算機ログ解析を支援するログ情報ブラウザ，情報処理学会論文誌 Vol. 41, No. 12, pp. 3265-3275, 2000.

[3] 牧野泰光，WWW におけるオブジェクト参照予測による応答速度向上に関する研究，北陸先端化学技術大学院修士論文，1998. http://www.jaist.ac.jp/library/thesis/is-master-1999/paper/y-makino/paper.pdf