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

2005－CG－121 (7)
2005／11／18

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教育部

‡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，LAN内にあるホストのアクセス傾向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，複数ホストのデータを同時に表示することが困難であり，ホスト間のデータを比較するなどの解析に限界がある．本稿では複数ホストのデータを総覧表示することが可能な可視化手法を提案し，上記の問題を解決した．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インタラクティブにソートやフィルタリングができる機能を提供している．

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 access pattern from each internal host. By using existing tools, it is difficult to compare to several host data because the log data size is too large to be displayed on the computer screen at the same time. Our visualization technique enables to display the several host data and to compare to a number of data. For visual data mining, there are several interactive operations we also provided.

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，匿名性確保などをおこなうことができる．これらのプロキシサーバの特性から，高速化やセキュリティの向上を目的として，ある程度の規模のネットワークではWebプロキシサーバを導入するケースが多い．Webプロキシサーバを設置しているネットワークの管理者は，効果的なキャッシングやフィルタリングをおこなうため，サーバ上のアクセスログを解析する必要がある．また，アクセスログを解析することによって，LAN内のインターネットユーザのアクセス動向を把握することも可能である．

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，アクセスログから必要な情報をわかりやすく提示するツールの需要は高く，アクセスログの解析ツールを開発してサービスをおこなっている企業もある[1] [2]．これらのサービスによるレポートは詳細であり膨大なログデータを扱えるが，結果表示は単純なグラフやテキストの羅列がほとんどである．そのため，レポートから有効な情報を得るのに手間がかかってしまうケースも少なくない．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ため，表示方法に工夫が必要となる．

そこで，本稿ではWebプロキシサーバのアクセスログを解析する際の手間を減らすこと，また，予期していない有用な情報が得られるビジュアルデータマイニングとしての可視化を目的とした可視化手法を提案する．

2. 既存手法

本節では，プロキシサーバのアクセスログの構造，既存手法を紹介する．

2.1 Webプロキシサーバのアクセスログ

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Webプロキシサーバであるsquid[3]のアクセスログの例である．アクセスログにはユーザがインターネットにアクセスした時刻，送信元ホストのIPアドレス，アクセス先のURL，送信サイズ，ファイル形式などのデータが記されておりWebページにアクセスする度にログが追加される．アクセスログを見れば，いつどのホストがどのWebページにアクセスしたかがわかる．Web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．大規模なログファイルでは一日あたり100MBを越える場合もある．

```
1125760014.117    61 *.*.11.82 TCP_MISS/200 45338 GET
アクセスした時間  送信元IPアドレス          送信サイズ

http://www.yahoo.co.jp/ - DIRECT/202.93.91.215 text/html
    送信先URL                                ファイル形式
```

図1：Webプロキシサーバ(squid)のログの例

2.2 既存手法

ログの解析は，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，可視化手法も変わる．既存のログ解析の可視化手法として，アクセス動向を把握するための可視化[4]やセキュリティの観点からの不審なログパターンを見つけるための可視化[5]が提案されている．

3. 提案手法

本節では，LAN内における複数ユーザのアクセス動向(ユーザがどのようなサイトにアクセスしているか，どのユーザがどれくらいの量のアクセスをしているか)がわかるような解析を目的とした可視化システムを提案する．

3.1 可視化手法

本手法では表状にデータを並べ，データを見比べることができるようにする．縦列に全ユーザのアクセスを合計した送信先URLをアクセス数が多い順に配置する．横列には各ユーザを左からアクセス数の合計が多い順に配置し，ユーザの各URLの全アクセス数における割合を表示する．全アクセス数における割合は少ない場合(0～25%)を青色，多い場合(75～100%)を赤色，その中間(25～75%)を緑色で表示する（図2）．この表示方法により，アクセスが多いサイトとそのサイトへの全ユーザのアクセス数が大まかに把握できる．また，そのサイトが多くのユーザからアクセスされている，単一のユーザからのみアクセスされているといったことも直感的に読み取ることが可能である．ユーザのアクセス動向を把握すること

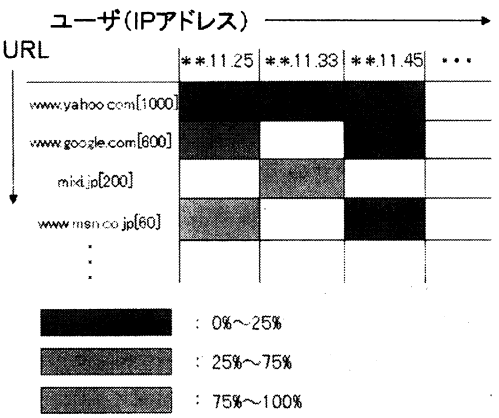


図 2：可視化手法概要

により，必要に応じてキャッシュ容量の見直しや好ましくないページをフィルタリングする際の判断材料になる．

3.2 インタラクティブ操作

解析するデータが複雑な場合，コンピュータが自動的にユーザの欲しい情報のみを表示することは難しい．そのような場合，有益な情報がありそうなポイントを人間が推測し，インタラクティブな操作によってデータを発見するビジュアルデータマイニングが効果的である．ビジュアルデータマイニングのメリットはユーザがデータ発見の過程に携われることである．文献 [6] で記述されているようにビジュアルデータマイニングは複雑なデータから有益な情報を得る場合，自動アルゴリズムよりも早くかつ良い結果が得られることが多い．

本システムではインタラクティブな操作として可視化データのソートとフィルタリングを用意する．

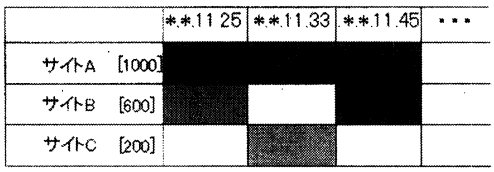
3.2.1 ソート

デフォルトの結果画面において，アクセス数の多い順に URL は表示されている．しかし，これらの順番をアクセス数以外の規則によってソートすることで，新しい有益な情報が得られる可能性がある．本システムでは，総アクセス数におけるユーザごとのアクセスの割合と特定のサイトにおけるアクセスの割合を比較し，差分の値によってソートすることも可能にした．これはユーザによってアクセスの絶対量が異なり，特定のサイトにおけるアクセスの割合を見たとき，総アクセス量が多いユーザの割合が大きいのは当然である．逆に，総アクセス量が少ないユーザの割合が大きい場合，そのサイトはより特徴的であると考えられる．したがって，このソートをおこなうことで特徴的なサイトを発見しやすくなることが期待される．

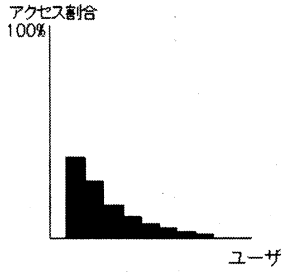
このソートを適用した例を図 3 に示す．図 3 (a) はソート前のデータ列である．サイト A，サイト B，サイト C といったサイトがあるとし，アクセス数が多いものから昇順で並んでいる．図 3 (b) は総アクセス数におけるユーザごとのアクセスの割合をグラフ化したものであり，図 3 (c)，図 3 (d) はサイト A とサイト C におけるユーザごとのアクセスの割合をグラフ化したものであるとする．図 3 (c) と図 3 (d) を比較すると，サイト A のパターンの方がより全体のパターンに類似しており，差分をとるとサイト C よりも小さくなる．このような方法で，差分が大きい方から昇順にソートすると図 3 (e) のようなデータ列となる．

3.2.2 フィルタリング

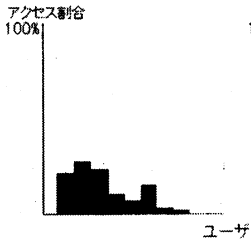
大規模なログを集計して得られた結果もまた膨大であることが多いため，そのままではまだ解析が困難であることが多い．そこで，有益な情報を得るためにデータを絞り込むフィルタリングが有効となる．フィルタリングには様々な方法が考えられるが，本システムでは URL にアクセスしたユーザの数で絞り込む．見つけたいデータの特徴をあらかじめ把握している場合，データを絞り込むことによって，有益な情報に近づく可能性が高くなる．しかし，有益な情報も排除してしまう可能性があるので注意する必要がある．



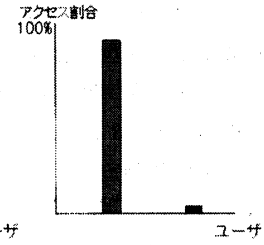
(a) : ソート前のデータ列



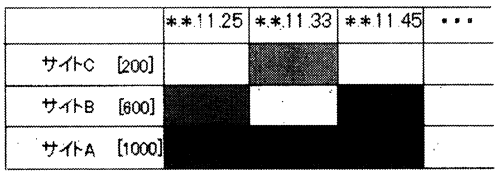
(b) : 全体のパターン



(c) : サイトAのパターン



(d) : サイトCのパターン



(e) : ソート後のデータ列

図 3 : ソート方法の具体例

4. 実験

本実験では, ある大規模ネットワークにおける Web プロキシサーバのログを用いる. Web プロキシサーバは squid を使用しており, ログは 2005 年のある一日の 0 時から 24 時の間に出力されたものである. ログファイルのサイズは 66MB であり, 約 44 万行のテキストが記されている. このログを本手法に適用した結果のデフォルト画面を図 4 に示す.

本手法を適用することによって, あるサイトに対しての全ユーザのアクセス状況を総覧することができる. これにより, 多くのユーザがアクセスしているサイトや, どのユーザのアクセスが多いかが把握できる. アクセス数が上位で多くのユーザからアクセスがあるサイト (青色が多く表示されているサイト) は YAHOO などのポータルサイトや有名なサイトである可能性が高く, 逆に少ないユーザからアクセスがあるサイト (赤色が含まれているサイト) は特徴的なサイトである可能性が高い. 図 4 の画面においても, 上位には多くのユーザからアクセスがあるサイトが見られ, その中に混じって赤色が描画されているサイトは注目するべきだと考えられる.

このような解析は, 1 画面に複数のユーザのアクセスデータを同時に表示することによって容易にできる. これまでの Web プロキシサーバのログ解析ツールでは, 1 ユーザごとの詳細なレポートが 1 画面ずつ表示されるものがほとんどである. そのような表示方法では, 複数ユーザのデータを比較することは困難であり, 本手法のような解析をおこなうには相当の手間が必要となる.

図 4 では, 全ユーザの総アクセス数の多い方から昇順に URL が表示されている. それによりアクセス数とアクセスユーザ数との関係を容易に見ることができ, 有益な情報を見つけるための判断材料となる. さらに別の判断材料として, 前章で説明したソートをおこなう. 図 4 のデフォルト画面をソートした結果画面を図 5 に示す. ソート後の結果は, アクセス数ではなく総アクセス数のパターンから外れている順に並んでおり, 異なる観点から判断することができる. この結果において最上位から 3 つ連続で同じユーザからのアクセスが多いことがわかる. これらのサイトはいずれも中国のサイトであり, このユーザは中国人もしくは中国に関心が高いと考えられる.

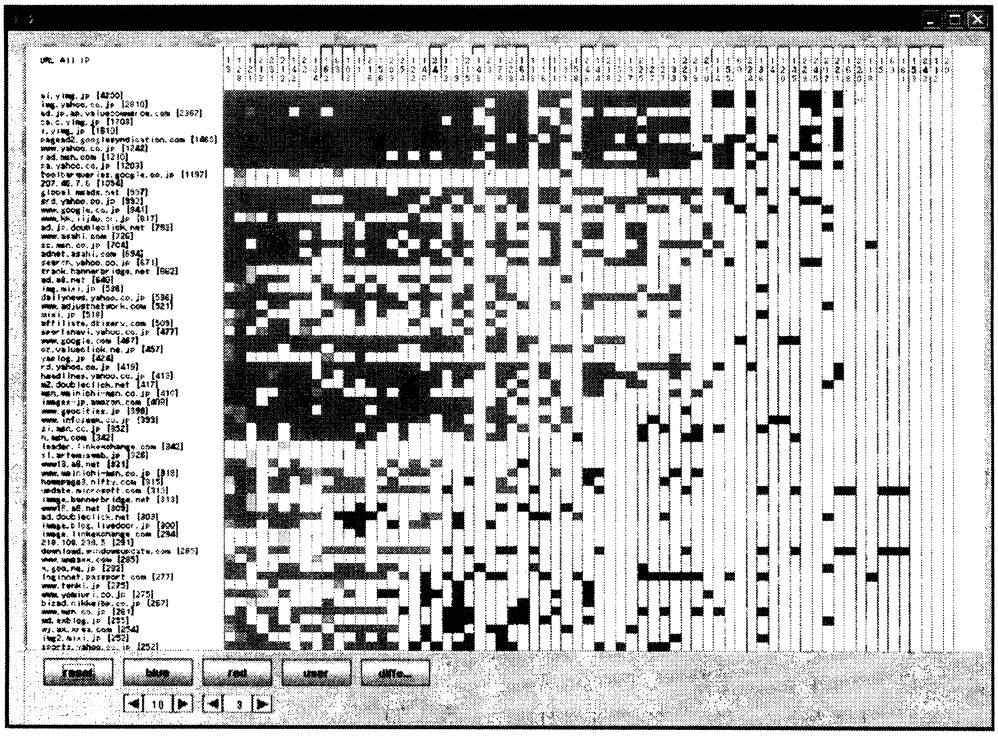


図 4：デフォルトの可視化画面（65 ユーザ）

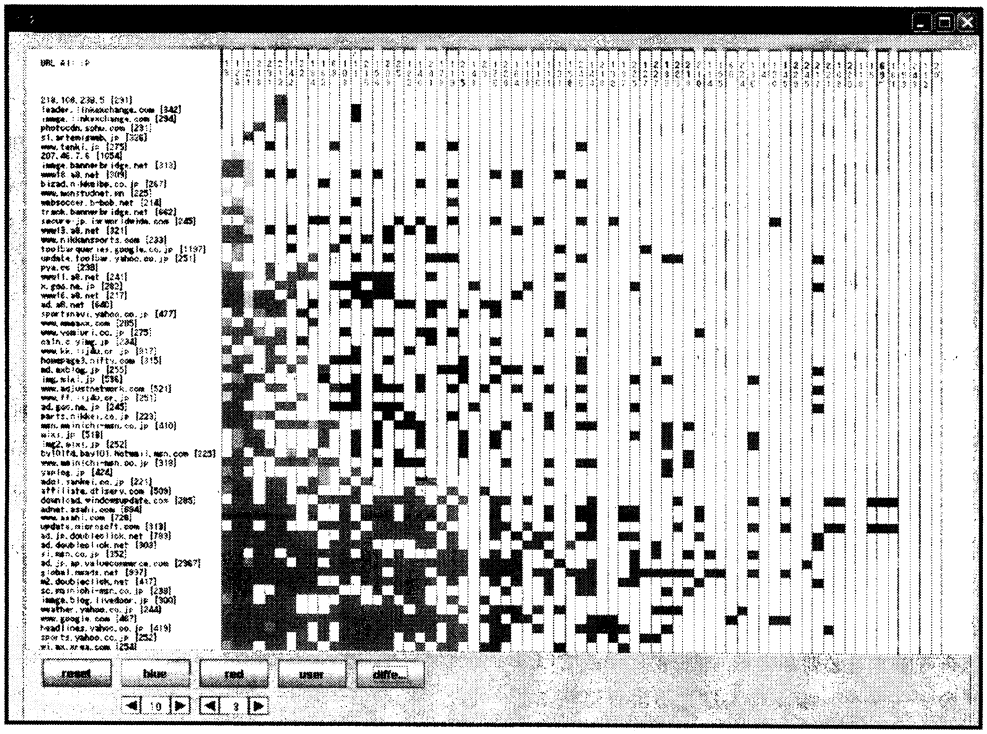
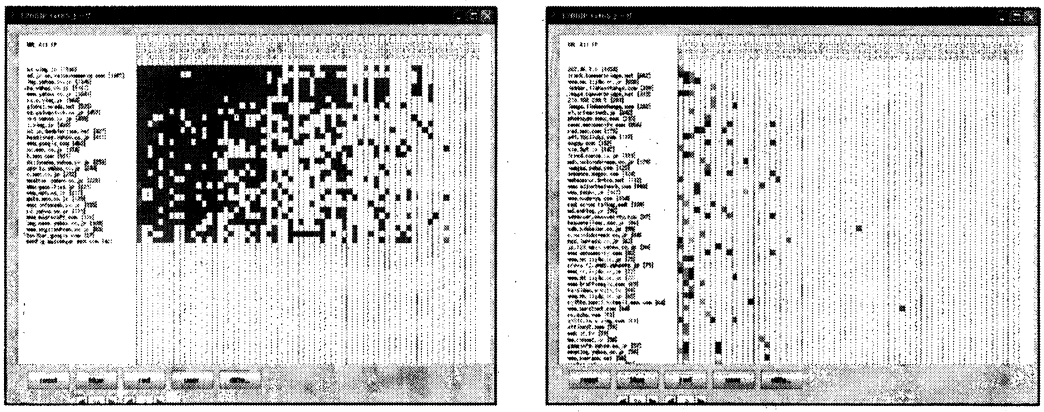


図 5：ソート後の可視化画面



(a) (b)

図 6：フィルタリング結果 (a) アクセスユーザ数 20 以上のみ表示，(b) アクセスユーザ数 3 以下のみ表示

図 6 にはフィルタリングをおこなった結果を示す．図 6 (a) はアクセスユーザ数が 20 以上のサイトのみを表示しており，図 6 (b) はアクセスユーザ数が 3 以下のサイトのみを表示している．フィルタリングをおこなうことによって，表示されるデータが少なくなり，より発見したいデータを見つけやすくなる．

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．一画面に全ユーザのデータを同時に表示することによって，既存のツールでは難しいアクセスユーザ数を考慮した解析を実現することができた．また，総アクセス数のパターンと各サイトのパターンを比較してのソートによる解析を提案した．

今後は時系列を考慮した可視化方法の提案を課題とする．例えば，一日のログデータにおいて何時に多くのユーザからのアクセスが偏っているなどが解析できると考えられる．長期間のデータを解析すれば人気が上昇しているサイトなどが見つけられる可能性がある．

また，ユーザ数がさらに多くなった場合の表示方法を考慮する必要がある．本可視化手法ではユーザのデータを一画面に入りきるように横に並べているが．この方法では，ユーザ数がある程度多くなった場合，ユーザのデータを表示するスペースが 1 ピクセルより小さくなってしまう可能性がある．

参考文献

[1] アクセス解析ソフトウェア サイトとラッカー，http://www.sitetracker.jp/tokucho.html#dorildownhttp://www.j-cyfin.com/.

[2]ログ解析ツール LogInspector (ログインスペクター)，https://loginspector.jp/accent/basic.html.

[3]Squid Web Proxy Cache，http://www.squid-cache.org/

[4] 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).

[5] 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).

[6]Daniel A.Keim，Information Visualization and Visual Data Mining，*IEEE Transactions on Visualization and Computer Graphics*, Vol.7，NO.1，pp.1-8(2002).