WWWトラフィック解析による情報生産者と情報消費者の数量的特徴付け

佐藤進也[†] 風間一洋[†]
清水奨[†] 神林隆[††]

WWWサーバ（httpdなど，WWWで情報リソースを提供するサーバ）とWWWにアクセスしているユーザのコミュニティの特徴をWWW proxyサーバのアクセス記録から導き出す方法について述べる．まず，サーバのランク（アクセス数による順位）と，当該サーバにアクセスしたクライアントの数という2つの数値の変動を，public proxyのログをもとに調べた結果を紹介する．特に，サーバを特徴付ける数量V（あるいは$V_{\log}$）を変動のスペクトルから導き出す方法と，その数量の基本的な性質を示す．また，コミュニティのWWW利用特性を把握する手段として，コミュニティがアクセスしたすべてのサーバについて$V_{\log}$を求め，その分布を調べるという方法を提案する．

Numerical Characterization of Information Producers and Consumers by WWW Traffic Analysis

SHIN-YA SATO,[†] KAZUHIRO KAZAMA,[†] SUSUMU SHIMIZU[†]
and TAKASHI KAMBAYASHI[††]

We describe methods to characterize a WWW server and a client community based on WWW access traces. We looked into access logs of a public WWW proxy, and found that a server can be characterized by a value V (or $V_{\log}$) derived from a spectrum of fluctuation patterns of a locus of (r, N_c), where r and N_c are respectively rank of the server and the number of clients that have accessed the server. We show the precise procedure to get V from the logs and its basic properties. We also introduce a method to characterize a client community, which is to see the distribution of $V_{\log}$ for all servers that the community has ever accessed.

1. はじめに

1.1 情報システムと社会性

近年，World Wide Web（WWW）という場を利用した情報の生産，消費活動が活発に行われている．さらに，企業などは商品の販売にも利用してきており，WWWは経済的な影響力も持ちつつある．計算機ネットワーク上で生まれたこの仮想情報空間は次第に我々の実生活に近い存在になってきている．その結果として，WWWも必然的に社会的な性質を帯びるようになり，そこに人間の活動パターンの一端をかいま見ることができる．たとえば，WWWのトラフィックを調べてみると，社会現象によく見られるZipfの法則[4),5)]や自己相似的特性[6)]が認められる．また，計算機を使った情報利用支援においても，社会的な概念の利用可能性が注目されている．協調情報フィルタリング（social filtering），協調型推薦システム（collaborative recommender systems）などがその例である[9)]．

1.2 広域分散情報検索システム

我々は，WWWにおける効率的な情報検索機構の構築を目指し，キーワード検索の手法をベースにした広域分散情報検索システム[7)]のプロトタイプを開発した．このシステムでは，検索サービスに必要なリソース（データを格納する記憶媒体，索引を作るための計算コストなど）を複数のサブシステムで分担することによりWWWにおける情報の増加に対応する．さらに，索引の作成などといった検索を制御する手段をも分散させて情報提供者に分け与え，情報の提供と検索という今まで分離されていた2つのプロセスをつなげる．これにより，情報をより効率的に流通させることを狙いとした．

† NTT未来ねっと研究所
NTT Network Innovation Laboratories
†† 日本テレマティーク株式会社
Nippon Telematique Inc.

1.3 新たなアプローチの必要性

さらに我々は，効率的な情報流通を支援するために，情報を参照するプロセス（情報消費のプロセス）を生産のプロセスにシームレスにつなげることを考えている．情報検索は情報の生産と消費をつなげる1つの方法である．たとえば，従来からよく用いられているキーワード検索では，生産された情報とその消費はキーワードによってつながりを得ることになる．しかし，キーワードが使われているコンテキストを生産者と消費者で共有することはできていない．その結果，検索者にとって不必要な情報が多く示されることは我々が検索サービスを利用しているときによく体験していることである．

キーワード検索の質を向上させる方法として検索者からのフィードバックを利用する方法などがすでに提案されている（たとえばSaltonら[1]の6章）．一方，計算の基本機能というレベルでは，情報の蓄積には安定したダイナミクスが適しているが，伝送や変換にはカオスなどの不安定なダイナミクスが適しているという知見が得られている[2]．これは，WWWという情報がアクティブに行き交う環境においては，従来の情報検索とは本質的に異なったアプローチが必要であることを示唆している．我々もまた，キーワード検索のような比較的静的な情報に対して有効な方法と相補的な役割を果たす，コンテキストをともなった情報の動き（変化）に適応できるシステム（アーキテクチャ）が必要であると考えている．

情報の動きを見るということは，人々の動き（挙動）を見ることである．ここでもまた，社会性を理解しそれを利用することが求められている．そもそも情報というものは人間が社会において活動している中で発生するものであるから，情報を扱ううえで社会性を考慮することは当然のことであるともいえよう．

1.4 情報利用活動の解析

そこで，我々は人間の情報利用活動の解析をはじめた．まず，いまWWWがどのように利用されているのか，ユーザが情報流通にどのように関与しているのかを調べるため，proxyサーバDeleGate[11]のログを解析した[8]．

ログに残されたデータの中で，我々は，WWWサーバ☆とクライアントの関係，すなわち，情報生産者と消費者の関係に注目した．具体的には，サーバのランク（アクセス数で決まる順位）と，そこにアクセスしたクライアントの数という2つの数値の変化を追った．その結果，変化のゆらぎから，WWWサーバのある種の状態（情報生産者としての活性度や成熟度など）を示していると思われる数量V☆☆を得た．本論文の目的は，この数量の性質を複数の側面から探り，その意味を帰納的に明らかにしていくことにある．

まず，Vを導く手法が一般的に適用可能であることを示す．先の解析[8]で対象としたアクセスログは，NTT研究所の一部という比較的小さい特殊な集団で使用しているproxyサーバのものであった．そこで，不特定多数のユーザが利用するpublic proxyのログを用いた解析を行った．その解析の過程と結果を次章以降で紹介する．

ところで，Vがサーバの状態を反映する数量である，という主張は，個々の事例を調べて得られた経験則に基づいている．この意味付けが，Vのもともとの定義と矛盾していないことを次に示す．Vの値を導き出すためにはサーバへのアクセス数を集計しランクを計算しなければならない．集計を開始する時点を変えると累積アクセス数が変化し，その結果，ランクの推移も変わる可能性がある．つまり，Vの値は見かけ上，集計開始時刻に依存している．しかし，Vがサーバの状態を表す数量であるならば，それは集計開始時刻に依存すべきではない．本論文では，この依存性が低いことを，集計開始時刻の変化にともなうVの値の変動幅の確率分布によって示す．

最後に，Vに期待されているもう1つの意味，すなわち，情報にアクセスしているユーザ集合（コミュニティ）の特徴付けについて議論する．あるサーバについて計算したVの値は，そのサーバの絶対的な評価ではなく，「あるコミュニティがそのサーバをどう見ているか」という相対的なものであると考えるべきである．本論文では，あるコミュニティがアクセスしたすべてのサーバに対してV（実際には，Vと同等な意味を持つ$V_{\log}$）を計算し，その分布によってコミュニティの特徴付けを行う，という方法を提案する．

1.5 本論文の構成

以降，2章ではVを導出する手順とすでに得られている解析結果を大まかに紹介する．3章において今回解析の対象としたデータの概要やデータ処理の過程を示す．さらに，Vを導く手法がpublic proxyのログデータに対しても適用でき，ローカルなproxyを対象にした場合と同様な結果が得られることを示す．4章ではVの基本的な性質を調べ，その中で集計開始

☆ httpd, ftpdなど，ネットワーク上で情報リソースを提供するサーバの総称．

☆☆ 'vitality' のV．

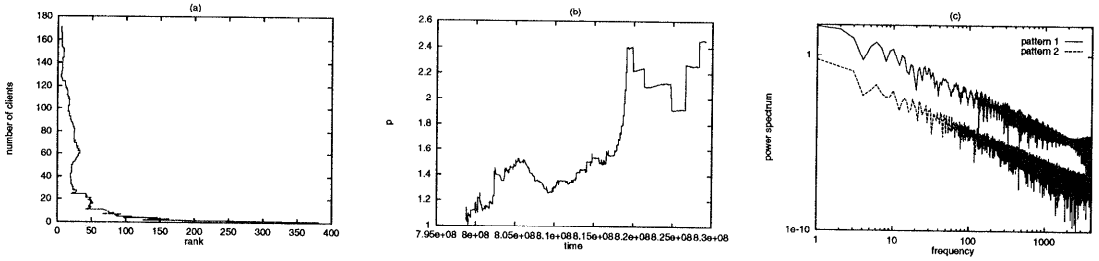


図1 一次解析の結果
Fig. 1 Preliminary results from the previous analysis.

時刻への依存性を調べる．5章ではコミュニティ特性としての V の分布について議論し，6章でその応用についてふれる．

2. 一次解析の概要と結果

本章以降の理解の助けとするため，ここでは，我々が以前行ったWWWアクセス履歴解析[8]の過程と得られた結果の概要を紹介する．

アクセス履歴はNTTの研究所で運用されていたproxyサーバDeleGate[11]のログから得た．このログに記録されていた1995年4月から約1年間，200万弱のアクセスのうち，イメージデータなどへのアクセスを除いて得られた70万弱のアクセスを解析対象とした．

あるWWWサーバに対して，ある時刻までにアクセスしてきたクライアントの数を N_c，延べアクセス数によって決まる人気順位（ランク）を r とすると，点 (r, N_c) の軌跡は図1(a)のように，$N_c = \alpha/r^p$ のグラフに似た形状を示した．

この式の α として，その時点までにログに現れたサーバの数をとったときの p の値を計算すると，図1(b)のような特徴的な変動のパターン（初期に見られる小刻みなパターン，後期に見られる階段状パターンなど）が認められた．p の変動のパワースペクトルを計算したところ，いずれも $1/f^2$ ゆらぎに近い特性を示しており，パターンの違いはスペクトルの強さの違いとして現れていることが確認できた（図1(c)）．

パワースペクトルが両対数グラフにおいて直線的な分布を示すことから，スペクトルの強さを求めることは，対数をとったスペクトルデータの分布から得られる回帰直線方程式の定数項を求めることに帰着できる．

以上の手順により，ログデータから自動的に，p のスペクトルの強さを計算できる．このようにして得られた値を V と定義した．安定した人気を保っているサーバでは V が高い値を示すことが経験的に分かっている．このため，V は情報生産者としての活性度あるいは成熟度といった状態を示しており，アクセス数とは違った側面からサーバを特徴付ける数量であると思われる．

また，あきらかに，情報参照者の集団に依存する数量であり，その意味においては，情報参照者の特徴をも表している数量である．

3. public proxy ログデータの解析

前章で紹介したデータ処理手法を不特定多数のユーザが利用するpublic proxyサーバのログに適用した結果を紹介する．データを処理した手順を詳しく示すとともに，以降で必要になる用語や記号を定義する．

解析に用いたのは，IMnetで公に提供されているキャッシュサービスのログ中1997年6月30日，4時01分37秒 ($= T_{start}$) から1997年11月30日，4時01分01秒 ($= T_{end}$) までの記録である．以降，時刻はシステム暦時間☆，すなわち1970年1月1日，00:00:00 UTCからの閏秒を除いた秒数で表すことにする．たとえば，$T_{start} = 867610897$，$T_{end} = 880830061$ である．

3.1 クライアントによるアクセスの抽出

このキャッシュサービスはSquid[10]を使っている．Squidでは，より効率的なproxyサービスを提供するため，複数のproxyサーバ間でキャッシュ情報を交換する．サーバのログにはこのキャッシュ情報交換の記録も含まれている．その中から，クライアントからの情報リソース（URLで指し示されHTTPなどのプロトコルで取得可能な文書や画像データなどの総称．以降リソースと略す）取得要求の記録だけを取り出した．その結果，18,522クライアント☆☆からの63,183,877件のリクエスト（リソース取得要求）を得た．これは，キャッシュ情報交換を含む全記録の31.2%にあたる．

☆ C言語のtime関数などの返り値に用いられている時間の表し方．

☆☆ 下位のproxyサーバもこれに含まれる．

3.2 クライアントのアドレスから名前への変換

ログ中には，クライアントを識別する情報としてIPアドレスのみが記録されている．後の処理のため，DNSの逆索きにより，名前（fully-qualified domain name; FQDN）に変換した．逆索きに失敗したものは解析対象から除外した．その結果18,522中17,376クライアントが残った．

3.3 Content typeによる随伴アクセスの除外

インラインイメージへのアクセスなど，情報への意図的アクセスに随伴して発生するアクセスを除外するため，解析対象をリソースのContent-Typeが`text/*`（subtypeは任意）であるものに限った．ログにContent-Typeが記録されていないものに関しては，HTMLファイルか検索の問合せと推測できるもの，具体的には要求しているリソースのURLが`.html`，`.htm`，`/`のいずれかで終わるか，あるいはURL中に`?`という文字が含まれているものを解析対象とした．

3.4 クライアントコミュニティの選択

今回の解析では，クライアントの集合（クライアントコミュニティ，誤解の恐れがない場合にはコミュニティと略す）を定義し，コミュニティごとに，その要素からのアクセスに限定した集計を行った．特に，任意のドメイン dom に対して，コミュニティ $C_{<dom>}$ を dom ドメインに属するホストの集合と定義する．また，ドメイン dom の1つ下のレベルのサブドメインの集合を $*.dom$ と書くことにする．たとえば，`nic.ad.jp` は `*.ad.jp` の要素である．今回，主に解析の対象としたのは，$C_{<\texttt{jp}>}$，$C_{<\texttt{ac.jp}>}$，$C_{<\texttt{or.jp}>}$ および，$\{C_{<i>}\}_{i\in *.\texttt{ac.jp}}$，$\{C_{<i>}\}_{i\in *.\texttt{or.jp}}$ の各コミュニティからのアクセスである☆．

いくつかのドメイン名と，それらに付随するクライアントコミュニティの要素数，ログ中に記録されていた当該コミュニティからのアクセス数（サーバの数とリソースの数），さらに下位ドメインの数を**表1**に示す．集合 A に対して A の要素数を $|A|$ で表す．

ここで，サーバ数はURL中のホスト名で相異なるもの，リソース数は相異なるURLの数を集計したものである．よって，同一ホストで複数のサービスを提供している場合（たとえばhttpとftp）でも，1つのサーバとしてカウントしている☆☆．なお，$|C_{<\texttt{jp}>}|$ がログに記録され名前が判明したクライアントの数17,376より少なくなっているのは`jp`ドメイン以外からもこのサーバが利用されているためであり，たとえば`net`ドメインからは全体の2%ほどの利用がある．

☆ 本論文では詳細は省略したが，$C_{<\texttt{ne.jp}>}$，$C_{<\texttt{ad.jp}>}$ などについても同様な結果が得られている．

☆☆ このようなサーバは全体の0.9%弱であった．

表1 クライアントコミュニティの例
Table 1 Some examples of client communities.

dom	$\lvert C_{<dom>}\rvert$	サーバ数	リソース数	$\lvert *.dom\rvert$
jp	15,823	104,801	2,958,644	14
ac.jp	1,781	83,750	2,187,685	146
or.jp	8,899	22,308	333,307	167

3.5 集計期間の決定と p の変動の計算

コミュニティ C から，サーバへのアクセス数などをカウントするにあたり，集計期間（T_c），すなわち集計の開始と終了時刻（$T_c = [t_0, t_1]$）を決めなければならない．今回，基本的には，ログに記録されているすべてのデータを集計対象としたので，$t_0 = T_{start}$，$t_1 = T_{end}$ である．

コミュニティ C，サーバ s，時刻 $t \in T_c$ に対して，t_0 から t までに C からアクセスされたサーバの集合を $S(C, T_c, t)$，アクセス数に基づく s の S における順位（ランク）を $r(C, T_c, s, t)$ とする．さらに，同じ期間に s にアクセスしてきたクライアントの数を $N_c(C, T_c, s, t)$ とする．二次元平面上の点 (r, N_c) の軌跡は，ローカルなproxyサーバのログを調べたときと同様に，$y = \alpha/x^p$ という巾関数のグラフに似た形を呈することを確認した．そこで，2章で紹介した方法に従い，p を次式が成り立つような関数として定義する．

$$N_c(C, T_c, s, t) = |S(C, T_c, t)|\, r(C, T_c, s, t)^{-p(C, T_c, s, t)}$$

これを，p について解くと

$$p(C, T_c, s, t) = \frac{\log |S(C, T_c, t)| - \log N_c(C, T_c, s, t)}{\log\ r(C, T_c, s, t)}$$

となる．この式が $r = 1$ のときにでも定義されるように，分母を補正したものが次式である．

$$p(C, T_c, s, t) = \frac{\log |S(C, T_c, t)| - \log N_c(C, T_c, s, t)}{\log\ r(C, T_c, s, t)\ + 1}$$

上式で定義された p の時間の推移に沿った変動を調べると，やはり図1(b)に見られるような小刻みなパターン，階段状パターンなどが認められた．なお，クライアント数 N_c の代わりにアクセス数を使った場合には，このような特徴的なパターンは見られない．

3.6 パワースペクトルと観測条件 M

次に，p の変動パターンの特徴を数量化するため，まず，そのパワースペクトルを計算する．具体的には，データのサンプリング，FFTによる周波数成分の取り出し，パワースペクトル計算という手順をふむ．サン

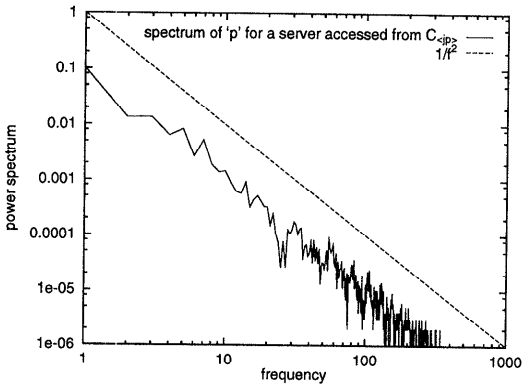


図 2 p のゆらぎのスペクトル
Fig. 2 Power spectrum of the fluctuation of p.

プリングの仕方を決めるのは次の3つのファクタである．1) 開始時刻 T_M（システム暦時間），2) 間隔 I_M（秒），3) サンプル数 N_M（個）．これらファクタの組 $M = \{T_M, I_M, N_M\}$ を観測条件と呼ぶことにする．集計期間を $[t_0, t_1]$ との間には，明らかに $t_0 \leq T_M$，$T_M + I_M N_M \leq t_1$ が成り立たなければならない．

図 2 は，$C_{<\text{jp}>}$ をクライアントコミュニティとするあるサーバのスペクトルを，$T_c = [T_{start}, T_{end}]$，$M = \{878000000, 300, 8192\}$ という条件で計算した結果である（両対数）．やはり，$1/f^2$ に近い特性を示しているのが分かる．なお，サンプル数は，FFT で処理するデータの数は2の巾でなければならないという条件と，適度に細かいサンプリング間隔（この場合5分）に対して，集計期間がパターンの特徴をとらえるのに適当な時間幅（約1カ月）になることを考慮して 8,192 とした．

3.7 回帰直線の計算

パワースペクトルを両対数グラフ上で見ると直線的分布を示しているので，対数をとったスペクトルデータの分布を回帰直線に帰着させ，その方程式により p の変動を特徴付ける．

得られた回帰直線の方程式のうち，回帰係数はほとんどサーバに依存しないため，結果的には方程式の定数項にサーバの特徴が現れていることになる．この定数項を $V_{\log}$ とし，V を $V = e^{V_{\log}}$ と定義する．V は，両対数グラフにおいて回帰直線が縦軸と交わる点でのスペクトルの値である．$V_{\log}$（あるいは V）は，クライアントコミュニティ C，サーバ s，集計期間 T_c，サンプリング条件 M を決めたうえで計算できる数量であるから，正確には，$V_{\log}(C, s, T_c, M)$ と書くべきである．

4. V の基本的性質

4.1 集計期間に対する依存性

サーバとそこにアクセスするコミュニティのある時点における特徴を反映した数量として $V_{\log}(C, s, T_c, M)$ がコミュニティ C，サーバ s，サンプリング M に依存するのは理にかなっている．しかし，$T_c = [t_0, t_1]$ の関与（実質的には t_0 の関与）を積極的に許す理由はない．ここでは，V の T_c，すなわち集計開始時刻 t_0 への依存性の有無を調べ，V の意味を検証する．

具体的には，2つの集計期間 $T_{c_0} = [t_{0_0}, T_{end}]$，$T_{c_1} = [t_{0_1}, T_{end}]$ と，適当なクライアントコミュニティの集合 $\mathbf{C}$ とサーバの集合 $\mathbf{S}$ に対して

$$\Delta = |V_{\log}(C, s, T_{c_0}, M) - V_{\log}(C, s, T_{c_1}, M)|, \quad C \in \mathbf{C},\ s \in \mathbf{S}$$

を計算し，その確率（頻度）分布を調べた．実際に計算に用いたコミュニティの集合は $\{C_{<d>}\}_{d \in *.\text{ac.jp}}$ であり，このコミュニティからアクセスのあったサーバの集合を $\mathbf{S}$ とした．また，$M = \{877979739, 300, 8192\}$ である☆．$t_{0_0} = 870203739$ とし，$t_{0_1} = 872795739$，875387739 の場合を比較対象とした．その結果を**図 3** (a) に示す．いずれの場合にも，$\Delta \leq 2.0$ という，$V_{\log}$ の分布（図 3 (b)）に比べて狭い範囲に9割以上が分布している

よって，Δ の値，すなわち2つの $V_{\log}$ の差は有意に小さいといえる．これは，$V_{\log}(C, s, T_c, M)$ の T_c に対する依存性が低いことを示している．

4.2 クライアントの多重化

今回のデータ解析は，多様で多数の proxy 利用者のアクセス履歴を対象としている点で以前の解析[8]と大きく異なっている．利用形態も様々で，たとえば proxy サーバを多段に使用している場合もあるだろう．このとき，1つのクライアント（の名前）が，複数のクライアントからのアクセスを代行しているという状況，すなわち，クライアントの多重化が発生する．多重化が本解析に与える影響を調べるため，proxy 経由のアクセスをシミュレートし，$V_{\log}$ の値を proxy を経由しない場合のものと比較した．

2つのクライアントコミュニティにおいて $C_1 \subset C_0$ が成り立っているとき，C_1 を C_0 のサブコミュニティと呼ぶことにする．たとえば，$C_{<\text{ac.jp}>}$ は $C_{<\text{jp}>}$ のサブコミュニティである．クライアントコミュニティ C_0 とそのサブコミュニティ C_1 を選び，ログ中の C_1

☆ 決められたサンプル数が得られないなど，$V_{\log}$ が正常に計算できない場合は調査対象から外した．

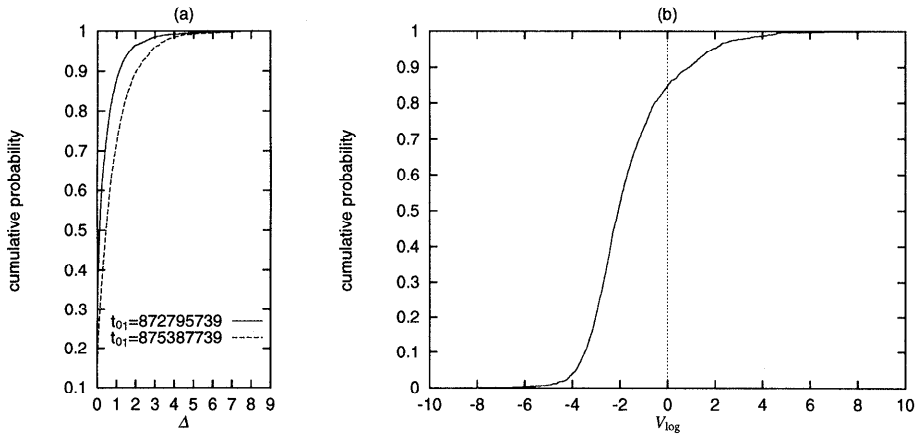


図 3 Δ と $V_{\log}$ の値の分布
Fig. 3 Probability distribution of Δ and $V_{\log}$.

表 2 シミュレーションに用いたクライアントコミュニティ
Table 2 Client communities analyzed in the simulations.

dom	$\lvert C_{<dom>}\rvert$	アクセス数
ac.jp	768	1,572,510
x.ac.jp	156	83,609
y.*x*.ac.jp	99	28,039
or.jp	2,544	184,947
z.or.jp	290	23,333

	C_0	C_1
(a)	$C_{<ac.jp>}$	$C_{<x.ac.jp>}$
(b)	$C_{<or.jp>}$	$C_{<z.or.jp>}$
(c)	$C_{<x.ac.jp>}$	$C_{<y.x.ac.jp>}$
(d)	$C_{<x.ac.jp>}$	$\{C_{<i>}\}$, $i \in *.x.ac.jp$

からのアクセスは要求 URL 単位に集約することで，proxy 経由のアクセスをシミュレートした．シミュレーションに用いたコミュニティとその特徴を**表 2** 左に示す．ここで，観測条件は $\{872795739, 300, 8192\}$ とし，$\lvert C_{<dom>}\rvert$（クライアント数），アクセス数はともにサンプリング時間区間内で集計したものである．また，*x*.ac.jp, *y*.*x*.ac.jp, *z*.or.jp はそれぞれ *.ac.jp, *.*x*.ac.jp, *.or.jp の中でクライアント数が最大のコミュニティを選んだものである．

シミュレーションは，表 2 右に示した (a)〜(d) の組合せで行った．各コミュニティからアクセスのあったサーバごとに，$V_{\log}$ の値を proxy のある/なしで比較した結果，その差が最大で 1.49（(a) の場合）〜8.42（(c) の場合）となった．proxy 経由のアクセスは $V_{\log}$ の値に影響を与えること，言い換えると，$V_{\log}$ の値は観測点（どの proxy サーバのログで計算するか）に依存することが分かった．なお，(d) は *.*x*.ac.jp の要素ごとに個別の proxy を利用した場合をシミュレートしたものである．

一方，$V_{\log}$ の分布は proxy の影響を受け難く比較的安定しているという結果を得ている（5.2.2 項）．

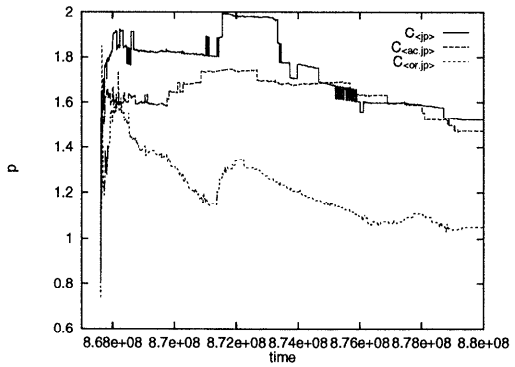


図 4 サブコミュニティと *p* の変動パターン
Fig. 4 Fluctuation patterns of *p* for subcommunities.

5. コミュニティの特徴付け

5.1 サブコミュニティと *p* の変動パターン

V の値のもととなる *p* の変動パターンはクライアントコミュニティによって異なる．ここでは特に，コミュニティとそのサブコミュニティの間でさえもパターンに本質的な違いが見られることを示す．

図 4 は，$C_{<jp>}$ とそのサブコミュニティ $C_{<ac.jp>}$, $C_{<or.jp>}$ からあるサーバへのアクセスを測って得られた *p* の変動を示すグラフである．$C_{<jp>}$ と $C_{<ac.jp>}$

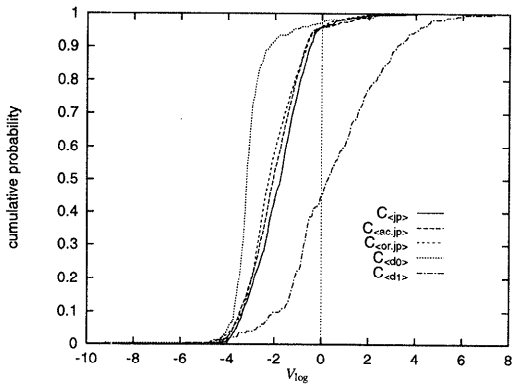


図5 V_{log} の値の分布
Fig. 5 Probability distribution of V_{log}.

のパターンは双方とも階段状を示している点で似ているが，$C_{<or.jp>}$ のパターンは $C_{<jp>}$ のものと本質的に異なっている．

5.2 V_{log} の分布

前節で確認したように，あるサーバ s_0 の見え方は異なるコミュニティ C_0，C_1 によっては違う場合がある．しかし，その事実からただちにコミュニティの情報アクセスのスタイル☆に違いがあると結論付けることはできない．たとえば，任意の s_0 に対してもう1つのサーバ s_1 があって，C_0 と s_0 との関係が C_1 と s_1 との関係と同様であり，かつ，C_0 と s_1 との関係が C_1 と s_0 との関係と同様であれば，全体として C_0 と C_1 は似た特性を持つと考えるべきだろう．そこで，コミュニティとしての特徴を知る一手段として，V_{log} の分布を見るという方法を提案する．なお，以下 V_{log} を V に置き換えても，値の範囲などが変わるだけで本質的には同様な議論が成り立つ．

5.2.1 分布の特徴

$C_{<jp>}$，$C_{<ac.jp>}$，$C_{<or.jp>}$，$C_{<d_0>}$，$C_{<d_1>}$ (d_0, $d_1 \in$ (*.ac.jp∪*.or.jp)) という5つのコミュニティの V_{log} の確率分布を図5に示す．

前節で見たように，$C_{<jp>}$ と $C_{<or.jp>}$ の間に局所的な違い（あるサーバの見え方の違い）があったが，分布の様子が似ていることから，似た特性を持ったコミュニティであると思われる．一方，$C_{<d_0>}$ や $C_{<d_1>}$ は $C_{<jp>}$ とは違った分布を示している．

一般に，データの分布のばらつきを表す数量として分散 σ がある．図5の $C_{<d_0>}$ のグラフで示されるような分布の分散は小さく，$C_{<d_1>}$ のような分布の分散は大きい，という傾向がある．この経験則に基づくと，分散から V_{log} の分布の大まかな特徴を推測できる．表3に C と V_{log} の分布の分散，さらに V_{log} の最小，最大値をまとめておく．

☆ アクセスの量の多少，頻度，あるいはサーバの利用度分布など情報アクセス形態に関する種々の要素を包含した抽象的概念を指す．

表3 コミュニティと V_{log} の分散
Table 3 Variances of V_{log} for some communities.

community	σ	min	max
$C_{<jp>}$	1.159479	−4.774582	4.127841
$C_{<ac.jp>}$	1.186232	−5.222968	3.302961
$C_{<or.jp>}$	1.159038	−6.715656	3.666590
$C_{<d_0>}$	1.046299	−9.150755	2.953076
$C_{<d_1>}$	2.098206	−5.714863	7.376623

5.2.2 クライアントの多重化と V_{log} の分布

クライアントの多重化シミュレーション（4.2節表2の(a)〜(d)）を使って V_{log} の分布に対する多重化の影響を調べた．その結果を図6に示す．多重化という変化に対して分布は安定していることが分かる．

5.2.3 コミュニティの和と V_{log} の分布

クライアントコミュニティ C_0，C_1 の和 C_0+C_1 を $C_0 \cup C_1$ で定義する．図7は，和という操作がコミュニティの V_{log} の分布にどのような影響を与えるかを，いくつかの特徴的な例によって示したものである．

図7(a)では，分布の“中和”が起きているのが見てとれる．一方，図7(b)では，やはり C_0+C_1 の分布が C_0，C_1 の分布の中間に位置しているものの，ほとんど片方（C_0）の分布と一致している．このような状況は，2つのコミュニティ間に情報アクセスの勢い（アクセス数，アクセス頻度）に大きな差がある場合によく見られる．しかし，情報アクセスの勢いだけが分布を決めるわけではない．(c)の例では，C_0 のアクセス数は C_1 の20倍以上にものぼる．それにもかかわらず，和の分布は C_0 によって一方的に支配されていない．さらに，和の分布が C_0，C_1 の分布の中間に位置していない点が特徴的である．

6. 考　　察

6.1 コミュニティの特徴付けについて

p の変動パターン，あるいは V_{log} の値は，コミュニティの変化に対して敏感であった．クライアントの多重化により V_{log} の値は変化し（4.2節），サブコミュニティにおいて p の変動パターンは変化した（5.1節）．それに比べて V_{log} の分布はそれらの変化に対して安定しており，コミュニティのマクロな特徴を反映していると思われる．たとえば，$C_{<jp>}$，$C_{<or.jp>}$，および $C_{<ac.jp>}$ の V_{log} の分布が似ているという事実（5.2.1項）は，この3つのコミュニティをマクロな観点で比較したとき似ていることを示唆している．実際この3

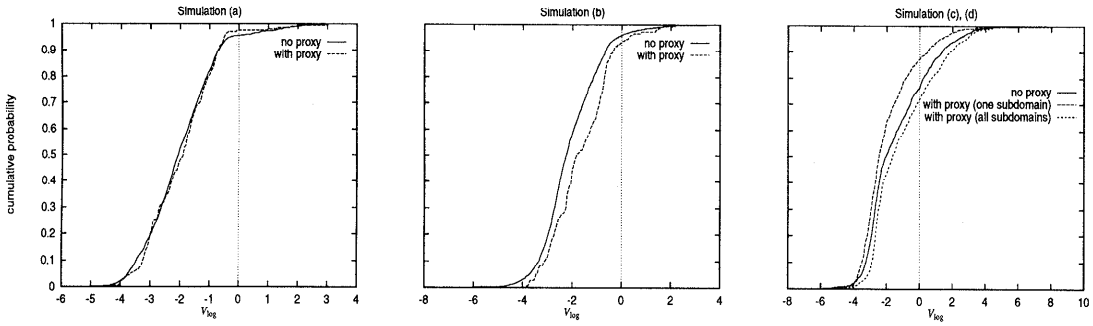


図 6 クライアントの多重化と $V_{\log}$ の値の分布
Fig. 6 Client multiplexing and $V_{\log}$ distributions.

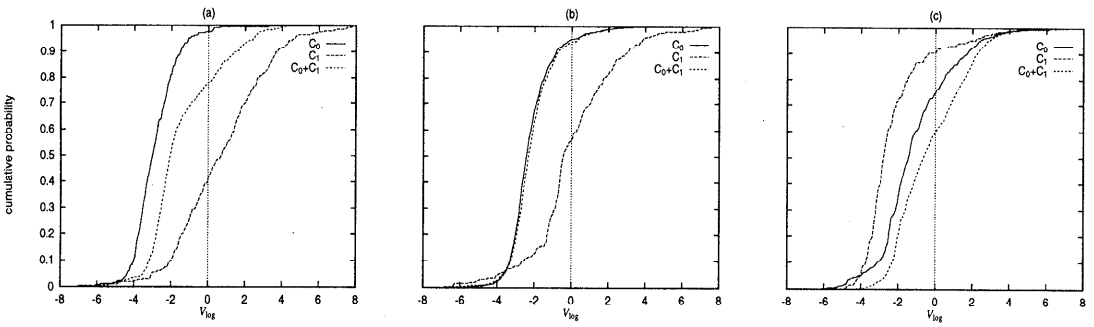


図 7 コミュニティの和と $V_{\log}$ の値の分布
Fig. 7 $V_{\log}$ for unions of communities.

者には，規模が大きく多くのサブコミュニティを含むという共通した特徴がある．

この $V_{\log}$ の分布を調べるという方法の特徴は，対象（情報生産者）を評価していることが結果的には自分（情報消費者）の評価になっている，という関係にある．同様な関係は実社会においてもみられる．たとえば，企業などの生産者によって市場に出された商品は消費者によって評価される．一方，消費者は逆に，その評価によって特徴付けられる．

さて，コミュニティを特徴付けるものとしては，メンバの数や延べアクセス数といった数量がある．これらのコミュニティの和 C_0+C_1 に対する値は，C_0，C_1 それぞれの値の和となり線形に変化する．一方，$V_{\log}$ の分布の変化は非線形である．2つのコミュニティの和に対して，それらの間をとったものとなっていることもあれば，ほとんど片方に一致すること，あるいは双方とも異なる分布になることもある（5.2.3 項）ため，コミュニティを構成する要素の加減がもとのコミュニティになかった新たな特性をもたらす可能性がある．

6.2 情報ブローカとしてのコミュニティ

$V_{\log}$ の意味を理解するにはさらに詳しい解析が必要であるが，その結果は，知的な情報ブローカの実現に寄与すると考えている．

情報ブローカとは，情報生産者と消費者の間に立って情報を中継し効率的な情報流通を助けるものである．たとえば，ディレクトリサービスも情報ブローカの一種である．情報ブローカは人間社会において実際に存在し，ブローカが存在することで情報流通が安定するという分析結果が報告されている[12]．

我々は，proxy サーバを利用するコミュニティで検索および閲覧の履歴を共有し，後の検索に利用する方法を考案し実装した[13]．これはコミュニティによって情報ブローカを実現しようとするものであり，その意味において，協調型推薦システム[9]と同じアプローチである．

このアプローチでは，コミュニティが変わるとブローカの特性も変わる．たとえば，多種多様なユーザのコミュニティで実現されている情報ブローカは大衆受けする情報を推薦するだろう．ある分野に特化した情報をブローカに推薦して欲しいのなら，ある条件に合致したユーザを選ぶといったコミュニティのチューニングが必要になる．チューニングの際に，$V_{\log}$ の分布によるコミュニティの特徴付けが一助となりうると考えている．

7. おわりに

WWWサーバのランクと，そこにアクセスしたクライアントの数という2つの数値の変動を，public proxyのログを対象として調べた結果を紹介した．ローカルなproxyのログを対象とした解析と同様に，変動は$1/f^2$ゆらぎを呈していることを確認し，そのスペクトルからサーバを特徴付ける数量Vを得た．Vは，情報生産者としての活性度あるいは成熟度といったサーバの状態を示していることが経験的に分かっていた．その解釈を支持する事実として，集計開始時刻に対してVの依存性が低いことを示した．ただし，この数量はクライアントコミュニティに依存する．よって，Vによって推測されるサーバの状態は絶対的なものではなく，クライアントコミュニティによる評価ととらえるべきである．

また，クライアントの集合であるコミュニティの情報消費活動の特徴を把握する手段として，コミュニティがアクセスしたすべてのサーバについて$V_{\log}$を求め，その分布を調べるという方法を提案した．$V_{\log}$の分布は，コミュニティの微小な変化に対して安定であり，コミュニティのマクロな特徴を反映していると思われる．

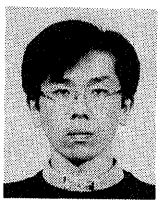
我々は，ここで得られた結果が,「情報流通においてコミュニティ形成を促すファクタとは何か」といった問いへの答えを探す手がかりとなると考えている．

謝辞 IMnet上のWWW public proxyの運用管理およびログデータの提供に対して，NTT情報流通プラットフォーム研究所の鍋島公章氏に感謝する．また，査読委員より詳細なコメントを頂戴した．特にクライアントの多重化に関しては，照会に基づいて大幅に改訂したことを付記して感謝したい．

参考文献

1) Salton, G., MacGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
2) 合原一幸，吉川敏文：秩序・カオス系と情報処理，人工知能学会誌，Vol.8, No.2, pp.179–183, 人工知能学会 (1993).
3) Zipf, G.K.: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Massachusetts (1949).
4) Cunha, C.R., Bestavros, A. and Crovella, M.E.: Characteristics of WWW Client-based Traces, Technical Report, BU-CS-95-010, Boston University Computer Science Department, Boston (1995).
5) Glassman, S.: A Caching Relay for the World Wide Web, *Computer Networks and ISDN Systems*, Vol.27, No.2, pp.165–173, Elsevier (1994).
6) Feldmann, A., Gilbert, A.C., Willinger, W. and Kurtz, T.G: Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic, *Proc. 35th Allerton Conference on Communication, Control and Computing*, Allerton House, Monticello, Illinois (1997).
7) Francis, P., Kambayashi, T., Sato, S., and Shimizu, S.: Ingrid: A Self-Configuring Information Navigation Infrastructure, *Proc. 4th International World Wide Web Conference*, Boston Massachusetts, World Wide Web Consortium (1995).
8) 佐藤進也，風間一洋，清水 奨：アクセス履歴を利用したWebサーバの状態の推定，*Proc. Japan World Wide Web Conference '97*, 横浜，日本インターネット協会 (1997).
9) Resnick, P. and Varian, H.R. (Eds.): Special Section: Recommender Systems, *Comm. ACM*, Vol.40, No.3, pp.56–89, ACM (1997).
10) Wessels, D.: Squid Internet Object Cache, <URL:http://squid.nlanr.net/>.
11) Sato, Y. and Hamazaki Y.: DeleGate: A General Purpose Application Level Gateway, *Proc. Worldwide Computing and Its Applications '97*, Lecture Notes in Computer Science, Vol.1274, pp.426–441, Springer (1997).
12) 後藤滋樹，野島久雄：人間社会の情報流通における三段構造の分析，人工知能学会誌，Vol.8, No.3, pp.348–356, 人工知能学会 (1993).
13) 清水 奨，神林 隆，佐藤進也，風間一洋：グループ指向WWW検索アシスタントPA-searchの実現，*Proc. Japan World Wide Web Conference '97*, 横浜，日本インターネット協会 (1997).

(平成10年4月9日受付)
(平成11年5月7日採録)



佐藤 進也（正会員）

昭和63年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．現在NTT未来ねっと研究所主任研究員．協調作業における情報活用支援の研究に従事．電子情報通信学会，Internet Society，ACM各会員．

風間　一洋（正会員）

昭和63年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在NTT未来ねっと研究所主任研究員．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM各会員．

清水　　奨（正会員）

平成4年東京大学工学部機械情報工学科卒業．同年日本電信電話（株）入社．現在NTT未来ねっと研究所．情報システムの研究に従事．Internet Society，ソフトウェア科学会，ACM各会員．

神林　　隆（正会員）

平成元年慶応義塾大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．現在日本テレマティーク（株）に出向．