社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# Webトラヒックの地域的な傾向分析

上山　憲昭[†,††]　中野　雄介[†,††]　塩本　公平[††]
長谷川　剛[†††]　村田　正幸[†]　宮原　秀夫[†]

† 大阪大学大学院情報科学研究科　〒 565-0871 大阪府吹田市山田丘 1–5
†† 日本電信電話株式会社 NTT ネットワーク基盤技術研究所　〒 180–8585 東京都武蔵野市緑 3–9–11
††† 大阪大学サイバーメディアセンター　〒 560-0043 大阪府豊中市待兼山町 1–32
E-mail: †kamiyama.noriaki@ist.osaka-u.ac.jp

**あらまし**　近年の Web サイトは，Ajax 等によって動的に生成されたオブジェクトを様々な配信ホストから取得する傾向を強めており，Web サイト閲覧時に発生する通信パタンが複雑化している．オブジェクトの配信には CDN(contents delivery network) が用いられることが多いが，レスポンス時間を改善し，ネットワーク内のトラヒック量を抑えるには，Web サイトアクセス時に生じるトラヒックの通信構造に基づき，キャッシュへのオブジェクト展開や置換処理等のキャッシュ制御を適切に行う必要がある．そこで本稿では，PlanetLab を用いて世界の 12 の拠点から，アクセス頻度の多い約 1,000 の Web サイトにアクセスしたときに発生する通信パタンを測定し，サーバ距離，遅延時間，オブジェクト数といった各種特性値の地域的な傾向について分析する．また各 Web サイトを 12 の各アクセス地点における各特性値の傾向に基づきクラスタ分析し，オブジェクト種別や URL カテゴリによる通信特性の違いを明らかにし，効率的なキャッシュ制御法を提言する．
**キーワード**　Web，アクティブ測定，通信構造

# Analysis of Locality Tendency of Web Traffic

Noriaki KAMIYAMA[†,††], Yusuke NAKANO[†,††], Kohei SHIOMOTO[††],
Go HASEGAWA[†††], Masayuki MURATA[†], and Hideo MIYAHARA[†]

† Department of Information Science, Osaka University　1-5, Yamadaoka, Suita, Osaka 565-0871
†† NTT Network Technology Laboratories, NTT Corporation　3-9-11, Midori, Musashino, Tokyo 180-8585
††† Cybermedia Center, Osaka University　1-32, Machikaneyama, Toyonaka, Osaka 560-0043
E-mail: †kamiyama.noriaki@ist.osaka-u.ac.jp

**Abstract**　Modern web-browsing services contain a number of rich objects, which are dynamically produced by servers and client PCs at diverse locations. Consequently, we face complications in understanding the communication structure of traffic generated when accessing websites. To improve the response time and suppress the amount of traffic transferred in networks, we need to adequately control caches, i.e., the placement and replacement of objects, based on the communication structure of web traffic. In this paper, we measure the communication structure of traffic generated when accessing about 1,000 most popular websites from 12 locations in the world using the PlanetLab. We investigate the locality tendencies of various metrics, e.g., server distance, delay, and object count, and we apply the cluster analysis to web sites based on the locality tendencies for various metrics in each object type and URL category. As a result, we find an insight to effectively improve the user response time through the cache control methods based on the locality tendencies of web sites.
**Key words**　Web, active measurement, communication structure

## 1. はじめに

近年，インターネットのトラヒックの多くの部分を，Web サービスで用いられる HTTP トラヒックが占めている．例えば 2006 年〜2008 年の間に日米間のバックボーンリンクで測定されたトラヒックの分析によると，HTTP パケットが約 60%を占めている [6]．Web サービスがインターネットの主流サービスであることから，Web トラヒックを適切に制御することが，ユーザの体感品質を向上させ，ネットワーク資源の消費量を抑えるためには重要である．

従来の Web サイトは静的なテキストや画像といったオブジェクトがサーバに用意され，Web ブラウザは HTTP を用いてこれら静的オブジェクトを単にダウンロードして表示していた．しかし近年，クライアント PC からのリクエスト受信時に，サーブレットや JSP(Java server pages) のプログラムをサーバ側で実行するか，JavaScript で書かれた Ajax や DOM によるプログラムを HTML に埋め込みクライアント PC 側で実行することで生成される動的オブジェクトの割合が増加している [7]．また，広告を専用のサーバから取得するなど，各オブジェクトの配信元が多元化している．このように一つの Web サイトを構成するオブジェクトは複雑性を増している．一方で，67%のユーザは毎週のように Web 閲覧時の待ち時間の長さを感じており，17%のユーザは Web 閲覧時に最大でも 5 秒しか待てないという報告がなされており [10]，複雑性を増す Web トラヒックをいかにして効率的に配信するかが重要な課題となっている．

Web 閲覧時のユーザのレスポンス時間を低減する技術としては CDN(contents delivery network) が一般的であり [17] [22]，アクセス数上位 1,000 のサイトの中では 74%が CDN を利用している [17]．CDN は主に Akamai 等の CDN 事業者が運営し

てきたが，近年，Google 等の大規模コンテンツプロバイダや，AT&T 等の Tier-1 ISP が自身で CDN を運用するケースも増えてきている [13]．キャッシュを用いたオブジェクト配信を行う際には，オブジェクトをコピーするキャッシュサーバの選択や，キャッシュの容量が不足した場合の削除オブジェクトの選択を適切に行う必要がある．

JPEG，JavaScript といったオブジェクトの種別や，Sports，Business といった URL カテゴリによって，配信要求の発生パタンが異なることが予想され，オブジェクト種別や URL カテゴリの違いを意識したキャッシュ制御を行うことが有効と思われるが，既存の方式ではオブジェクト間の差別化がなされていない．限られたキャッシュサーバ資源を有効に活用し，Web アクセス時のレスポンス時間を効果的に低減するためには，キャッシュ展開が有効なオブジェクトの種別を把握することが重要である．そのため著者らは，測定用 PC から多数のサイトにアクセスした際に発生するトラヒックの通信特性値を HAR(HTTP Archive) ファイルとして取得し，HAR ファイルから各種特性値を抽出することで，URL カテゴリやオブジェクト種別ごとの各種通信特性の傾向について分析したが [16]，単一地点 (東京) からの測定分析に限定されており，通信構造の地理的な傾向の差異は分析されていない．

そこで本稿では，PlanetLab [18] を用いて世界の 12 の拠点から，アクセス頻度の高い約 1,000 の Web サイトにアクセスしたときに発生する通信パタンを測定し，サーバ距離，遅延時間，オブジェクト数といった各種特性値の地域的な傾向について分析する．また各 Web サイトを各測定地点における各特性値の傾向に基づきクラスタ分析し，オブジェクト種別や URL カテゴリによる通信特性の違いを明らかにし，効率的なキャッシュ制御法を提言する．

## 2. 関 連 研 究

アクティブ測定による Web トラヒック分析に関する研究としては，Baeza-Yates らの，2004 年前後に 12 か国で実施した Web クローリングの測定結果から，国ごとの Web 閲覧トラヒックのサイズ，接続グラフの次数等の各種傾向を比較した研究や [4]，Butkiewicz らの，ランダムに選択した Web サイトを 9 週間にわたり周期的にアクセスし，構成オブジェクト数やアクセスサーバ数等の指標について分析した研究が見られる [7]．

一方，パッシブ測定による Web トラヒック分析に関する研究としては，Ihm らの，2006 年〜2010 年の 5 年間にわたる Web の proxy access log を用いて Web トラヒックの各種指標の変移を分析した研究や [11]，Bent らの，2004 年の 1 日のパケットキャプチャデータから，Web サイトの Cookiee 利用頻度等について分析した研究が見られる [5]．また Gill らは，企業や大学からの Web アクセストラヒックを分析し，Web サービスの利用傾向について明らかにしており [8]，Ager らは CDN やデータセンタ上のコンテンツ配置や，配信元サーバの選択がどのように行われているかを，DNS に関する制御パケットの測定と，BGP routing table の snapshot に基づき識別することを検討している [1]．また Schneider らはパケットキャプチャデータから HTTP や AJAX セッションを抽出し，生成されるトラヒックパタンの差異を分析している [21]．しかしこれらの研究では，Web サイトにアクセスしたときの，オブジェクトの配信元サーバとクライアント PC 間の距離といった，地理的な通信構造については分析されていない．

## 3. 実 験 手 法

本節では，多地点からの Web 通信構造分析のための測定実験手順について述べる．測定手順は，(i)PlanetLab 上での測定環境の構築と測定地点の選択，(ii) 評価 URL リストの生成，(iii) 各測定地点から各評価 URL にアクセスしたときの HAR(HTTP Archive) ファイルの取得，(iv)HAR ファイルからのデータ抽出，(v)RTT の測定，の 5 つの手順で構成される．以下に，各々の手順について述べる．

### 3.1 PlanetLab 上での測定環境の構築と測定地点の選択

PlanetLab はインターネット上に構築されたオーバレイネットワークで，世界の様々な地域に存在する約 500 のノードから構成される．PlanetLab を用いることで，選択したノード上で様々なプログラムを実行することができる．そのため (ii) 以降の手順を PlanetLab 上の複数のノードで独立に実行することで，世界中の様々な地域から様々な Web サイトにアクセスし，通信特性の情報を収集する．実験に先立ち，PlanetLab 上での測定実験環境を構築する必要があるが，PlanetLab が提供する GUI を用いて測定に用いるノードを起動する．北米 (NA) から三つ，欧州 (EU) から二つ，ロシア (RU) から一つ，オセアニア (OA) から二つ，南米 (SA) から二つ，アジア (AS) から一つ，そしてアフリカ (AF) から一つの，合計で 12 の PlanetLab ノードを測定ホストとして選択した．これら 12 の測定地点を表 1 にまとめる．

表 1　Measurement locations

| ID | Area | Location | ID | Area | Location |
|---|---|---|---|---|---|
| L1 | NA | Massachusetts | L7 | OA | Australia |
| L2 | NA | Wisconsin | L8 | OA | New Zealand |
| L3 | NA | California | L9 | AS | Japan |
| L4 | EU | Ireland | L10 | SA | Ecuador |
| L5 | EU | Germany | L11 | SA | Argentina |
| L6 | RU | Russia | L12 | AF | Reunion |

表 2　Website count used in clustering analysis, average object size, object count, and total data size in each URL category

| ID | Category | Website count | | Object size | Object count | Total size |
|---|---|---|---|---|---|---|
| | | 0:00 | 12:00 | (kbytes) | | (Mbytes) |
| C1 | Business | 59 | 40 | 14.70 | 55.14 | 0.810 |
| C2 | Computers | 112 | 91 | 16.26 | 43.63 | 0.709 |
| C3 | News | 39 | 27 | 13.55 | 72.45 | 0.982 |
| C4 | Reference | 112 | 109 | 13.09 | 43.42 | 0.568 |
| C5 | Regional | 80 | 73 | 17.77 | 50.59 | 0.899 |
| C6 | Science | 95 | 86 | 14.04 | 52.86 | 0.742 |
| C7 | Society | 79 | 83 | 15.01 | 66.86 | 1.003 |
| C8 | Health | 86 | 52 | 14.27 | 54.30 | 0.775 |
| C9 | Home | 85 | 47 | 15.66 | 55.39 | 0.867 |
| C10 | Shopping | 69 | 68 | 15.67 | 70.77 | 1.109 |
| C11 | Adult | 112 | 102 | 10.49 | 53.04 | 0.557 |
| C12 | Arts | 55 | 60 | 15.43 | 68.18 | 1.052 |
| C13 | Games | 87 | 58 | 15.28 | 54.12 | 0.827 |
| C14 | Kids & teens | 106 | 64 | 13.23 | 54.59 | 0.722 |
| C15 | Recreation | 86 | 52 | 13.55 | 57.30 | 0.776 |
| C16 | Sports | 38 | 53 | 16.62 | 86.67 | 1.440 |

### 3.2 評価 URL リストの生成

Web 通信構造の傾向を分析するためには，アクセス数の多い，高人気のサイトにアクセスしたときに発生する通信を分析対象とすることが望ましい．そこで Alexa のサイト [2] 上で公開されているアクセスランキングを元に，16 の各 URL カテゴリから，最もアクセス数の多い上位 300 の Web サイトを測定対象として選択した．いくつかの Web サイトは複数のカテゴリに重複して分類されているため，重複したサイトを削除することで 4,290 の Web サイトを測定対象に選択した．ランキングリストに記載されている URL は全て各サイトのトップページであるため，トップページにアクセスしたときの通信特性のみが分析対象となる．

### 3.3 評価 URL アクセス時の HAR ファイルの取得

次に，測定対象 URL リストと，HAR ファイル取得や HAR ファイルから統計データを抽出するプログラムを，(i) で起動した PlanetLab ノードにアップロードする．Web サイトにアクセスする時間帯によって，生じる通信特性が異なることが予想されるため，様々な測定地点間で Web 通信構造の傾向を比較するためには，全ての測定地点において同一の現地時刻に開始する必要がある．そこで UNIX の cron コマンドを用いて，UTC(coordinated universal time) より取得した各測定地点の現地時刻が midnight(0:00) と noon(12:00) となるときに各々，各測定用 PlanetLab ノードから 4,290 の Web サイトに連続してアクセスする実験を開始した．12 の全ての測定地点において HAR ファイルが正しく取得されたサイト (midnight で 1,124，noon で 927) を最終的に分析対象とした．表 2 に，各 URL カテゴリの分析対象 Web サイト数をまとめる．

生成した評価 URL リストの各 URL に対して，測定用 PlanetLab ノードから GET の HTTP リクエストを送信した際に発生する通信特性を HAR ファイルとして取得した．HAR ファイルは，クライアント PC とサーバ間で転送される HTTP データのヘッダ情報から，クライアント PC において，各オブジェクトのサーバ URL，サイズ，取得に要した遅延時間等の各種通信特性を算出し，JSON(JavaScript Object Notation) 形式で出力したものである [15]．多数のサイトにアクセスするために，コマンドラインで JavaScript を実行できる phantomjs 上で動作するスクリプト netsniff.js [9] を用いることで，多数のサイトに連続してアクセスし，各々の HAR ファイルをバッチ処理で取得した．この際に，測定用 PlanetLab ノードのローカルキャッシュを無効化することで，全てのオブジェクトをリモートのサーバから取得した．

### 3.4 HAR ファイルからのデータ抽出

取得された各 HAR ファイルに含まれる各構成オブジェクトの情報から，分析に必要となるデータを各測定用 PlanetLab ノードにて抽出する．具体的には，ホスト名，ホストの位置，オブジェクトサイズ，オブジェクト取得遅延時間，オブジェクト種別 (MIME Type) に関する情報を取得するため，HAR ファイル中の各 key に対応する value を Python で抽出した．オブジェクト取得遅延時間は，各オブジェクトに対する Request が測定用ノードから送信開始された時刻から，そのオブジェクトの測定用ノードでの到着が完了した時刻までの経過時間である．なお，MaxMind の提供する GeoIP API [14] を用いて，URL からサーバの存在する国名・都市名・位置座標を取得し，各オブジェクトの配信元サーバの位置座標と測定用 PlanetLab ノード間のユークリッド距離をオブジェクト距離と定義して算出する．そして抽出された統計データを，東京に設置した収集用 PC に転送する．

### 3.5 RTT の測定

オブジェクト距離は測定ノードとオブジェクトサーバ間のユークリッド距離であり，インターネット上での実際の距離とは異なる．そこで前ステップで述べた総計データに加えて，PlanetLab の各測定ノードにおいて各 Web サイトにアクセスして HAR ファイルを取得した直後に，各オブジェクトサーバに対して測定ノードから ping コマンドを送ることで RTT(round-trip time) を測定する．取得した RTT のデータは同様に，収集用 PC に転送する．

## 4. 平均値特性

表 2 に各 URL カテゴリの Web サイトの，平均オブジェクトサイズ (kbytes)，オブジェクト数，総データサイズ (Mbytes) をまとめる．これらの値は 12 の全測定地点における平均値である．総データ量とオブジェクト数は，Arts，Shopping，Sport といった娯楽系サイトで多い傾向が見られるのに対して，Business，Computers，Health，Reference といった情報収集系サイトで少ない傾向が見られる．アクセス都市数，アクセスホスト数に関しても同様の傾向が得られた．オブジェクト種別 (MIME Type) ごとに平均オブジェクトサイズ (kbytes) を算出すると，JavaScript が 30.42，HTML が 17.50，CSS が 23.12，JPEG が 9.59，PNG が 5.33，GIF が 1.96，となった．予想に反して，画像オブジェクトはサイズが小さく，テキストオブジェクトは大きい傾向が見られる．多くの Web サイトでは，アイコンといったサイズの小さい画像オブジェクトが多数を占める一方，Ajax 等を用いた Web サイトのリッチ化に伴い，動的生成オブジェクトを生成するプログラムが大規模化していることが理由と思われる．

以下，各測定地点の現地時刻で 12:00(noon) に取得した 927 の Web サイトの統計データを用いて，12 の測定地点の通信特性を比較する．表 3 に，各 URL カテゴリと 927 の全測定サイトにおける平均オブジェクト距離を示す．北米，欧州，南米からは小さいが，オセアニア，アジアからは大きく，北米の 7 倍ほどになっている．また多くの地域において，Business と Sports の平均オブジェクト距離が小さな傾向が見られる．一方，表の掲載は省略するが，各 URL カテゴリのオブジェクトまでの平均 RTT を比較すると，北米，欧州，ロシア，アジアからは小さいが，オセアニアや南米からは北米の 4 倍程度に，アフリカからは北米の 7 倍程度となった．また，ほぼ全ての測定地点において Home と Shopping の平均 RTT が小さく，Reference や Adult の平均 RTT が大きい．また各 MIME Type の平均 RTT についても分析したところ，全 12 測定地点において JavaScript は小さいが PNG や CSS は大きい．

表 4 に各 URL カテゴリの各オブジェクト取得に要した平均遅延時間をまとめる．北米や欧州からは小さいが，ロシア，オセアニア，南米，アジアからは北米の 2 倍程度，アフリカからは北米の 5 倍程度と大きい．多くの地域において，Shopping と Recreation のサイトは平均遅延時間が小さいのに対して，Adult と Reference のサイトは大きな傾向が見られる．アフリカからアクセスしたときはカテゴリ間の差異が小さい．表 5 に同様に各 MIME Type における平均遅延時間を示すが，全地域において JavaScript と CSS の平均遅延時間が小さいのに対して，JPEG は大きい．表 2 で見たように，画像オブジェクトのサイズは小さい傾向が見られるが，一方で，JPEG オブジェクトを提供しているサーバは高負荷である傾向が予想される．

また表の掲載は省略するが，構成オブジェクト数，アクセスホスト数，アクセス都市数の平均値は，全測定地点において同様の値となった．Arts，News，Shopping，Society，Sports のサイトは平均アクセスホスト数が多く，Adult と Reference は小さい傾向が全地域において見られる．また Computers と Reference の平均オブジェクト数は少ない一方，Arts，News，Shopping，Society，Sports のサイトは全地域において多い傾向が見られる．さらに News，Shopping，Sports は平均アクセス都市数が多い一方，Adult と Reference は全地域で少ない．

## 5. Web サイトのクラスタ分析

様々な地域からアクセスされたときの各特性値の傾向の違いを明らかにするため，We サイトの各特性値の地理的傾向に基づいたクラスタ分析を行う．各測定時刻 $t$ において，$N$ 個の様々な特定地点 $X_1$，$X_2$，$\cdots X_N$ から，$M$ 個の様々な Web サイト $Y_1$，$Y_2$，$\cdots Y_M$ にアクセスしたとき，各 Web サイトに対して平均 RTT 等の各特性値の $N$ 個の結果が取得される．よって，時刻 $t$ に Web サイト $y$ に測定地点 $k$ よりアクセスしたときの特性値を $v_{y,t,k}$ $(1 \leqq k \leqq N)$ とするとき，$v_{y,t,k}$ を要素にもつ $N$ 次元のベクトル $\boldsymbol{v}(y,t)$ を構成することができ，$1 \leqq y \leqq M$ の各 $y$ に対して，$M$ 個のベクトル $\boldsymbol{v}(y,t)$ が得られる．各特性値 $v$ のアクセス地点ごとの傾向の違いを分析するために，得られた $M$ 個のベクトル $\boldsymbol{v}(y,t)$ を用いて，k-menas 法を用いてクラスタ分析を行う．

### 5.1 k-means 法の初期クラスタ生成法

k-means 法の分類結果は初期クラスタに強く依存するため，最適なクラスタ構成が生成されるよう初期クラスタを構成することが重要となるが，本稿では Arthur らによって提案された k-means++法を用いて初期クラスタを構成する [3]．k-means++法は，$k$ 個の初期クラスタの重心をできるだけ偏りなく散らばらせるものであり，最初に一つのメンバをクラスタ重心としてランダムに選択し，以後，残るメンバの各々の最近接クラスタ重心までの距離の自乗に比例する確率でランダムに一つのメンバをクラスタ重心として選択する処理を，$k$ 個のクラスタ重心が選択されるまで反復する．ところで各特性値 $v$ のクラスタ重心は，そのクラスタを構成する全てのメンバの $v$ の平均値ベクトルであり．各要素は，構成メンバの各測定地点における $v$ の平均値を表す．

### 5.2 クラスタ数 $k$ の最適化

k-means 法の結果はクラスタ数 $k$ にも強く依存するため，$k$ を適切に設定することが重要となる．本稿では，Jain らによって提案された JD 法を用いてパラメタ $k$ を最適設計する [12]．各クラスタに属するメンバとそのクラスタ重心間の距離を最小化するのと同時に，任意の二つのクラスタの重心間の距離を最大化するために，JD 法は次式で定義されるコスト関数 $p(k)$ を最小化する $k$ を選択する．

$$p(k) = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \leqq j \leqq k} \left\{ \frac{\eta_i + \eta_j}{\xi_{ij}} \right\}$$

ただし，$\eta_j = \sum_{i=1}^{n_j} D(\boldsymbol{x}_i^{(j)}, \boldsymbol{m}_j)/n_j$，$\xi_{ij} = D(\boldsymbol{m}_i - \boldsymbol{m}_j)$ であ

表 3　Average distance of objects in each URL category ($10^3$km)

| | NA | | | EU | | RU | OA | | AS | SA | | AF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 |
| adult | 3.30 | 2.75 | 3.41 | 7.09 | 8.32 | 10.23 | 21.10 | 23.58 | 5.59 | 9.21 | 18.50 | 13.36 |
| arts | 2.96 | 2.33 | 2.34 | 6.64 | 8.54 | 9.69 | 18.32 | 20.58 | 5.33 | 8.31 | 18.45 | 13.33 |
| business | 2.18 | 2.50 | 4.08 | 6.13 | 6.70 | 7.49 | 18.67 | 19.33 | 5.49 | 8.70 | 17.41 | 12.67 |
| computers | 3.04 | 2.54 | 2.39 | 6.95 | 8.33 | 9.56 | 20.05 | 22.16 | 5.38 | 9.03 | 18.34 | 13.29 |
| games | 3.74 | 3.14 | 2.86 | 6.75 | 8.16 | 9.81 | 19.77 | 22.52 | 5.60 | 9.15 | 18.33 | 13.13 |
| health | 2.84 | 2.11 | 2.40 | 7.61 | 9.64 | 11.16 | 21.27 | 23.71 | 4.87 | 8.70 | 20.05 | 14.19 |
| home | 2.56 | 2.00 | 2.43 | 7.23 | 8.96 | 10.66 | 19.98 | 22.21 | 4.72 | 8.57 | 19.22 | 13.39 |
| kids and teens | 3.43 | 2.78 | 2.96 | 7.16 | 8.86 | 10.44 | 20.11 | 22.13 | 5.38 | 9.04 | 18.85 | 13.75 |
| news | 2.85 | 2.31 | 2.71 | 6.35 | 8.36 | 9.52 | 18.76 | 22.20 | 5.21 | 8.64 | 18.07 | 13.31 |
| recreation | 3.11 | 2.49 | 2.67 | 6.56 | 8.07 | 9.02 | 18.50 | 19.67 | 5.56 | 8.52 | 16.94 | 13.30 |
| regional | 3.15 | 2.92 | 3.30 | 6.51 | 7.90 | 8.26 | 17.91 | 18.91 | 5.77 | 8.68 | 17.34 | 12.76 |
| reference | 3.58 | 3.20 | 3.66 | 6.63 | 7.96 | 9.33 | 19.51 | 21.96 | 6.05 | 9.36 | 17.83 | 12.71 |
| science | 3.28 | 2.78 | 3.12 | 6.71 | 8.33 | 9.72 | 19.49 | 21.81 | 5.31 | 9.03 | 17.93 | 13.18 |
| shopping | 2.25 | 2.01 | 3.07 | 6.12 | 8.02 | 7.95 | 17.73 | 18.97 | 4.95 | 8.13 | 18.11 | 13.10 |
| society | 3.17 | 2.78 | 3.17 | 6.38 | 8.20 | 8.93 | 18.12 | 20.38 | 5.59 | 8.78 | 17.35 | 13.04 |
| sports | 2.93 | 2.38 | 3.19 | 5.39 | 7.95 | 8.35 | 15.74 | 17.42 | 5.00 | 7.68 | 17.10 | 12.61 |
| all | 3.13 | 2.66 | 3.06 | 6.70 | 8.29 | 9.45 | 19.22 | 21.27 | 5.45 | 8.82 | 18.13 | 13.19 |

表 4　Average total delay of objects in each URL category (seconds)

| | NA | | | EU | | RU | OA | | AS | SA | | AF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 |
| adult | 0.24 | 0.33 | 0.34 | 0.45 | 0.50 | 0.56 | 0.84 | 0.91 | 0.60 | 0.80 | 0.76 | 1.40 |
| arts | 0.18 | 0.31 | 0.23 | 0.30 | 0.37 | 0.58 | 0.56 | 0.69 | 0.71 | 0.61 | 0.48 | 1.14 |
| business | 0.18 | 0.29 | 0.29 | 0.27 | 0.28 | 0.54 | 0.47 | 0.65 | 0.72 | 0.83 | 0.37 | 1.39 |
| computers | 0.16 | 0.30 | 0.22 | 0.28 | 0.38 | 0.60 | 0.44 | 0.58 | 0.67 | 0.60 | 0.37 | 1.10 |
| games | 0.21 | 0.33 | 0.26 | 0.32 | 0.39 | 0.57 | 0.61 | 0.73 | 0.85 | 0.66 | 0.51 | 1.24 |
| health | 0.21 | 0.31 | 0.27 | 0.37 | 0.44 | 0.67 | 0.66 | 0.76 | 0.91 | 0.68 | 0.57 | 1.20 |
| home | 0.15 | 0.29 | 0.22 | 0.30 | 0.32 | 0.65 | 0.52 | 0.65 | 0.73 | 0.55 | 0.47 | 1.12 |
| kids and teens | 0.22 | 0.31 | 0.26 | 0.31 | 0.36 | 0.61 | 0.64 | 0.65 | 0.72 | 0.62 | 0.50 | 1.22 |
| news | 0.21 | 0.50 | 0.25 | 0.38 | 0.53 | 0.61 | 0.63 | 0.74 | 0.57 | 0.70 | 0.50 | 1.21 |
| recreation | 0.16 | 0.25 | 0.26 | 0.25 | 0.30 | 0.38 | 0.53 | 0.62 | 0.46 | 0.62 | 0.43 | 1.11 |
| regional | 0.18 | 0.33 | 0.24 | 0.26 | 0.30 | 0.41 | 0.44 | 0.76 | 0.48 | 0.69 | 0.37 | 1.34 |
| reference | 0.28 | 0.39 | 0.35 | 0.44 | 0.45 | 0.56 | 0.75 | 0.78 | 0.84 | 0.85 | 0.59 | 1.32 |
| science | 0.20 | 0.29 | 0.25 | 0.28 | 0.30 | 0.47 | 0.58 | 0.64 | 0.59 | 0.63 | 0.46 | 1.07 |
| shopping | 0.15 | 0.25 | 0.19 | 0.25 | 0.30 | 0.42 | 0.41 | 0.58 | 0.47 | 0.69 | 0.32 | 1.26 |
| society | 0.21 | 0.39 | 0.27 | 0.32 | 0.39 | 0.54 | 0.59 | 0.71 | 0.58 | 0.74 | 0.50 | 1.24 |
| sports | 0.19 | 0.29 | 0.25 | 0.23 | 0.36 | 0.37 | 0.52 | 0.69 | 0.61 | 0.57 | 0.40 | 1.06 |
| all | 0.20 | 0.32 | 0.27 | 0.32 | 0.38 | 0.53 | 0.59 | 0.70 | 0.66 | 0.69 | 0.49 | 1.22 |

表 5　Average total delay of objects of each MIME type (seconds)

| | NA | | | EU | | RU | OA | | AS | SA | | AF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 |
| JPEG | 0.26 | 0.41 | 0.38 | 0.44 | 0.61 | 0.75 | 0.87 | 1.01 | 0.90 | 1.02 | 0.73 | 1.79 |
| HTML | 0.25 | 0.38 | 0.30 | 0.32 | 0.36 | 0.59 | 0.59 | 0.76 | 0.66 | 0.68 | 0.48 | 1.16 |
| PNG | 0.21 | 0.33 | 0.29 | 0.33 | 0.35 | 0.51 | 0.63 | 0.69 | 0.69 | 0.72 | 0.55 | 1.25 |
| CSS | 0.16 | 0.23 | 0.21 | 0.25 | 0.39 | 0.41 | 0.51 | 0.62 | 0.57 | 0.60 | 0.41 | 1.00 |
| GIF | 0.20 | 0.30 | 0.26 | 0.30 | 0.29 | 0.47 | 0.63 | 0.66 | 0.60 | 0.64 | 0.52 | 1.07 |
| JavaScript | 0.15 | 0.27 | 0.19 | 0.22 | 0.36 | 0.39 | 0.39 | 0.55 | 0.53 | 0.51 | 0.30 | 1.06 |
| all | 0.21 | 0.33 | 0.28 | 0.33 | 0.41 | 0.54 | 0.63 | 0.74 | 0.68 | 0.72 | 0.52 | 1.29 |

り，$n_j$ はクラスタ $j$ に分類されたメンバ数で，$D(\boldsymbol{a}-\boldsymbol{b})$ は二つのベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ との間の距離である．そして $1 \leqq k \leqq 1+\log_2 n$ の範囲で $p(k)$ を最小化する $k$ を選択する．

### 5.3　数 値 結 果

12 の特性値を用いて Web サイトをクラスタ分析したところ，表 6 に示す値が $k$ の最適値として選択された(注1)．多くの特性値に対して Web サイトは 3 程度の少数のクラスタに分類されたが，平均オブジェクト取得遅延時間やレスポンス時間に対しては多数のクラスタに分類されており，これら特性値の地域的な傾向が多様であることがわかる．図 1(a) に midnight のデータに対して，平均オブジェクト距離に基づきクラスタ分析を行ったときに，生成された各クラスタの重心を 12 の各測定地点に対してプロットする．また図 1(b) には midnight のデータを対象に，16 の各 URL カテゴリと全 Web サイト (all) において，各クラスタに分類された Web サイトの比率をプロットする．ただしクラスタ番号は意味をもたないため，ここでは構成 Web サイト数の降順にクラスタ番号を付与した．すなわちクラスタ 1 が最大クラスタとなる．図 1(c) と (d) には同様に，noon のデータを対象に同様の結果を示す．midnight の平均オブジェクト距離の傾向は noon の傾向とほとんど同じであることが確認できる(注2)．

図 1(b)(d) から確認されるように，約 80%の Web サイトはサイズの最も大きいクラスタ 1 とクラスタ 2 に分類されているが，クラスタ 1 は，北米においては近く，欧州，南米，ロシア，アフリカにおいては中程度で，オセアニアとアジアは遠い．そのため，大多数の配信サーバやキャッシュサーバは北米に設置されていることが予想される．またクラスタ 2 もクラスタ 1 と似た傾向があるが，より世界の様々な地域にサーバが分散されている傾向が見られる．Business，News，Shopping，Sports は，クラスタ 1 よりもクラスタ 2 により多くのサイトが分類されており，これらサイトのオブジェクトは北米に加えて，より様々な地域に存在する．Adult，Health，Society のサイトはクラスタ 1 に分類される比率が高く，これらオブジェクトを提供するサーバは，より北米に集中する傾向が見られる．

図 2 には同様に，オブジェクト距離の分散について Web サ

(注1)：クラスタ分析の効果を高めるため $3 \leqq k \leqq 1+\log_2 n$ の範囲で $k$ を選択した．ただし $k$ の上限 $\log_2 n$ は，midnight で 11，noon で 10 となる．

(注2)：12 の全特性値において時間帯による傾向の差異は小さい．

イトをクラスタ分析したときに構成された各クラスタの重心と，各々のクラスタに分類されたサイトの比率をプロットする．約半数の Web サイトが分類されたクラスタ 1 の重心は，全ての測定地点において小さいことから，多くのオブジェクトは近隣に存在するサーバから取得されることが確認できる．他の二つのクラスタにおいては，オセアニアとアジアにおいてオブジェクト距離の分散が大きく，これらクラスタに分類されたサイトのオブジェクトはオセアニア，アジアにおいてキャッシュされておらず，北米の様々な場所から取得されている．Adult と Reference のサイトはクラスタ 1 への分類比率が高く，近隣に存在するサーバからオブジェクトが取得される傾向が強い．他方，News や Sports のサイトはクラスタ 1 への分類比率が低く，これらサイトのオブジェクトは世界の様々な場所に広く存在する傾向が強い．

表 6　Optimum value of $k$ in k-means method

| Property | Midnight | Noon |
|---|---|---|
| Average distance of objects | 5 | 5 |
| Variance of distance of object | 3 | 3 |
| Average RTT of objects | 7 | 3 |
| Variance of RTT of objects | 4 | 3 |
| Average object size | 3 | 3 |
| Variance of object size | 3 | 3 |
| Average total delay of objects | 8 | 9 |
| Variance of total delay of objects | 4 | 5 |
| Number of objects | 5 | 3 |
| Number of hosts accessed | 3 | 5 |
| Number of cities accessed | 3 | 3 |
| Response time | 7 | 7 |

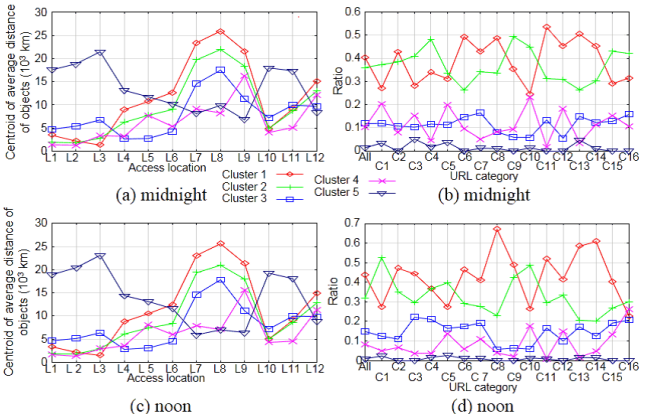


図 1　(a)(c) Centroid of average object distance at each access location, (b)(d) ratio of websites classified into each cluster in each URL category

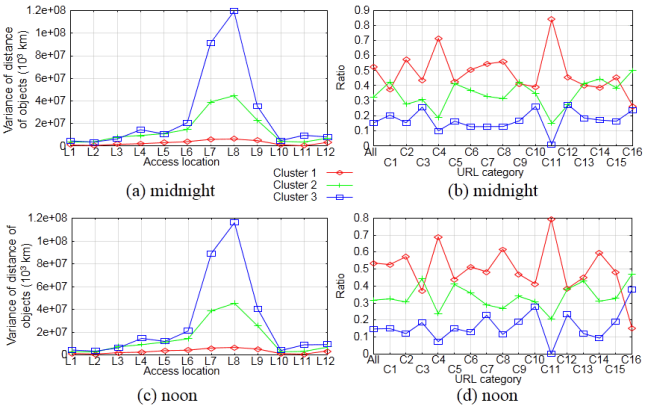


図 2　(a)(c) Centroid of variance of object distance at each access location, (b)(d) ratio of websites classified into each cluster in each URL category

図 3 に平均 RTT に対する生成クラスタの重心と分類比率を同様に示す．クラスタごとにより明確な地域的傾向の差異が確認できる．midnight におけるクラスタ 1 と 4，noon におけるクラスタ 1 は，南米とアフリカを除いて全ての地域で平均 RTT が小さい．midnight のクラスタ 3 と noon のクラスタ 2 は，北米において平均 RTT が小さいが，他の地域においては大きい．midnight のクラスタ 6 と 7，noon のクラスタ 3 は，欧州とロシアにおいて平均 RTT が小さく，北米を含め他の地域では大きい．Adult や Reference を提供するサーバの多くは北米に存在するため，北米以外の地域においては平均 RTT が大きい．

次に図 4 に，各オブジェクトの取得に要した平均遅延時間に関して，生成されたクラスタの重心と各カテゴリの各クラスタに対する分類比率をプロットする．多くのクラスタは似たような地理的傾向を示しており，各クラスタの重心の大小関係は，ほぼ全ての測定地点において同じになっている．例えばクラスタ 1 の重心は，12 の全ての測定地点において，他のクラスタの重心よりも小さく，平均遅延時間が小さい．また midnight のクラスタ 5 と noon のクラスタ 7 の重心は，全ての測定地点において他の多くのクラスタの重心よりも大きく，平均遅延時間が大きい．多くのクラスタの重心は，北米，欧州，アジアにおいて小さいが，他の地域では大きい．

また図 5 に，アクセス都市数に対して生成された各クラスタの重心と，各カテゴリの各クラスタへの分類比率を同様にプロットする．クラスタ 1 のサイズはクラスタ 2 のサイズとほぼ同じであることから，midnight においてクラスタ 1(クラスタ 2) に分類されたサイトは，noon においてはクラスタ 2(クラスタ 1) に分類されていることが予想される．ほぼ全てのサイトは，12 の全測定地点におけるアクセス都市数の差異が小さいが，クラスタ 3 に分類された約 10%のサイトは，欧州，ロシア，アフリカにおいてアクセス都市数が少ない．クラスタ 3 に分類された Adult，Computers，Reference のサイトは少なく，これらカテゴリのサイトを閲覧した際にアクセスされる都市数は少ない傾向が見られる．一方で，Business，News，Shopping，Sports のサイトはクラスタ 3 への分類比率が高く，これらサイトを閲覧した際には多数の都市にアクセスされる．図については省略するが，アクセスホスト数とオブジェクト数については，このような地理的傾向の差異は見られなかった．

## 6.　測定実験結果による主な知見

本節では，4. と 5. 節で述べた実験とクラスタ分析から得られた主な知見をまとめる．

**地理的傾向**

- 約 80%の Web サイトのオブジェクトは北米から配信され，約 15%のサイトのオブジェクトは欧州から配信され，そして数%のサイトはアジアから配信されている．そのため北米や欧州以外の地域からアクセスした場合，多くのサイトにおいて，オブジェクト距離，RTT，オブジェクト取得遅延時間が大きくなる傾向が見られる．
- 北米に多くのオブジェクト配信サーバが集中しているため，アジアやオセアニアからのオブジェクト距離は大きく，北米の 7 倍程度となる．しかし RTT は 4 倍程度，オブジェクト取得遅延時間は北米の 2 倍程度に留まっており，アジアやオセアニアから北米に至る NW の状態は良好である．
- アフリカからアクセスしたときは，平均オブジェクト距離，RTT，遅延時間が全て，他の地域と比較して突出して大きく，アフリカから北米に至る NW の状態は悪い．

**URL カテゴリの傾向**

- Business，News，Shopping，Sports といった，地域ごとに特有の情報が発信され，情報の地域性が高いカテゴリのサイトは，各々のアクセス地点の近隣に存在するサーバからオブジェクトが取得される．
- Health や Society といった，地域的な普遍性の高い情報を提供するサイトのオブジェクトの配信サーバは北米に集中しており，北米以外の地域からアクセスした際のオブジェクト取得遅延時間や RTT は大きくなる．この傾向は，特に Adult のサイトにおいて顕著である．
- Adult，Games，News，Reference といったサイトへのアクセス需要は，これらサイトのオブジェクトを提供する配信サーバの容量と比較して大きく，これら配信サーバやその周辺の NW は高負荷な場合が多く，これらサイトにアクセスしたときの RTT やオブジェクト取得遅延時間が大きな傾向が見られる．

**オブジェクト種別による傾向**

- 画像オブジェクトよりもテキストオブジェクトのサイズが大きい傾向が見られる．またテキストオブジェクトは遠方から，画像オブジェクトは近隣から取得される傾向が見られることから，画像オブジェクトはテキストオブジェクトと比較して地域性が高いことが予想される．
- 全ての地域において，JPEG オブジェクトの遅延時間の平均や分散は大きく，PNG オブジェクトの RTT は大きい．一方で，JavaScript オブジェクトの遅延時間や RTT の平均と分散は小さい．そのため，JavaScript オブジェクトの提供サーバや NW の状態が良好なのに対して，JPEG や PNG オブジェクトの提供サーバや NW は高負荷な傾向が予想される．

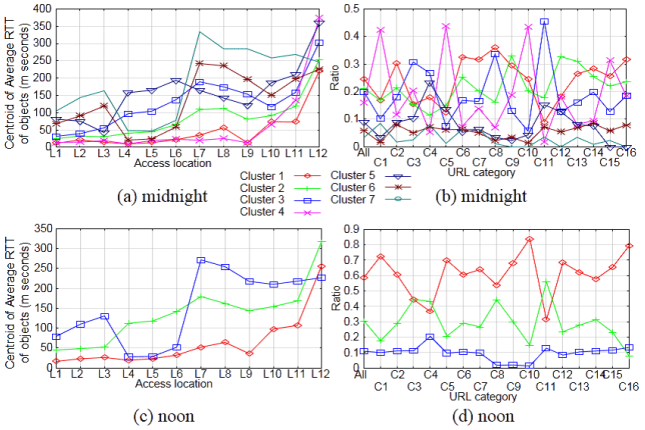


図 3 (a)(c) Centroid of average RTT of objects at each access location, (b)(d) ratio of websites classified into each cluster in each URL category

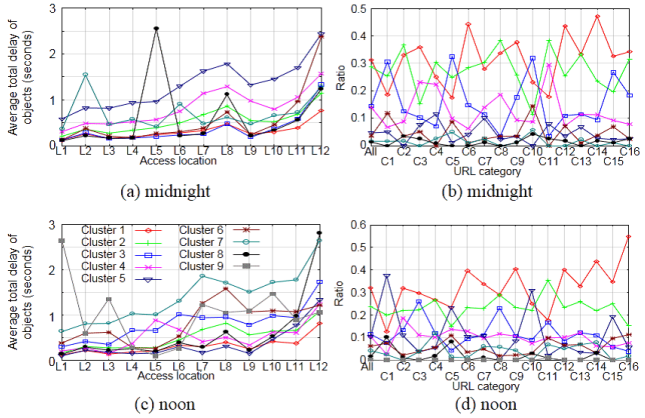


図 4 (a)(c) Centroid of average total delay of objects at each access location, (b)(d) ratio of websites classified into each cluster in each URL category

これらの知見をキャッシュ制御に生かすことが可能であり，例えば，JPEG や PNG オブジェクトや Adult や Reference といった情報の地域普遍性の高いサイトのオブジェクトを提供するサーバは高負荷で，北米に集中して配置されている傾向が見られることから，Web サイト閲覧時のレスポンスを改善するためには，これらのオブジェクトを北米に加えて様々な地域で優先的にキャッシュすることで，これらのオブジェクトを提供するサーバの負荷を低減することが有効である．

## 7. ま と め

Web サイトにアクセスした際のレスポンス時間を改善し，ネットワーク内のトラヒック量を抑えるには，Web サイトアクセス時に生じるトラヒックの通信構造に基づき，キャッシュへのオブジェクト展開や置換処理等のキャッシュ制御を適切に行う必要がある．そこで本稿では，PlanetLab を用いて世界の 12 の拠点から約 1,000 個の Web サイトにアクセスし，オブジェクト距離やオブジェクト取得遅延時間といった各種通信特性を測定し，平均値特性を比較することで，URL カテゴリやオブジェクト種別による通信特性の傾向の差異を分析した．また各 Web サイトを 12 のアクセス拠点における通信特性に基づきクラスタ分析を行い，カテゴリやオブジェクト種別による地域的な傾向の差異を分析した．

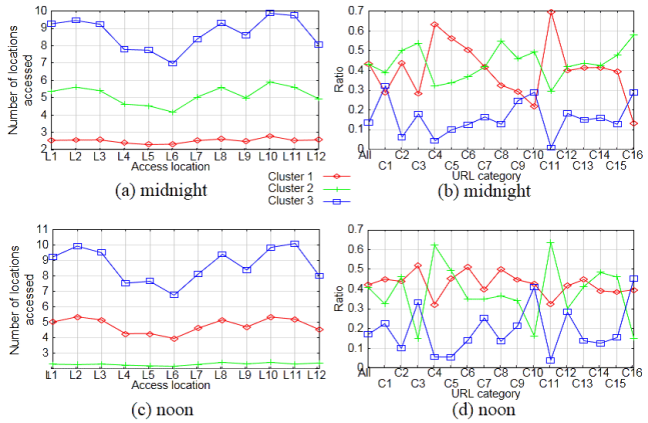


図 5 (a)(c) Centroid of number of locations accessed from each access location, (b)(d) ratio of websites classified into each cluster in each URL category

## 文 献


[1] B. Ager, W. Muhlbauer, G. Smaragdakis, and S. Uhlig, Web Content Cartography, ACM IMC 2011.
[2] Alexa, http://www.alexa.com/topsites/category.
[3] D. Arthur and S. Vassilvitskii, k-means++: the advantages of careful seeding, ACM SODA 2007.
[4] R. Baeza-Yates, C. Castillo, E. N. Efthimiadis, Characterization of national Web domains, ACM Trans. Internet Technology (TOIT), 7(2), Article No.9, 2007.
[5] L. Bent, M. Rabinovich, G. M. Voelker, Z. Xiao, Characterization of a Large Web Site Population with Implications for Content Delivery, ACM WWW 2004.
[6] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, Seven Years and One Day: Sketching the Evolution of Internet Traffic, IEEE INFOCOM 2009.
[7] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, Understanding Website Complexity: Measurements, Metrics, and Implications, ACM IMC 2011
[8] P. Gill, M. Arlitt, N. Carlsson, and A. Mahanti, Characterizing Organizational Use of Web-based Services: Methodology, Challenges, Observations, and Insights, ACM Trans. The Web, 5(4), Article No. 19, 2011.
[9] GitHub, Network Monitoring, http://github.com/ariya/phantomjs/wiki/Network-Monitoring
[10] When seconds count. http://www.gomez.com/wp-content/downloads/GomezWebSpeedSurvey.pdf.
[11] S. Ihm and V. Pal, Towards Understanding Modern Web Traffic, ACM IMC 2011.
[12] A. L. Jain and R. C. Dubes, Algorithms for clustering data, Englewood Clis, NJ Prentice-Hall, 1988.
[13] C. Labovitz, S. Iekel-Johnson, J. Oberheide, and F. Jahanian, Internet Inter-Domain Traffic, ACM SIGCOMM 2010.
[14] MaxMind, GeoIP Downloadable Databases, http://dev.maxmind.com/geoip/downloadable.
[15] J. Odvarko, HAR Viewer, Software is hard, http://www.softwareishard.com/blog/har-viewer.
[16] 上山, 中野, 塩本, アクティブ測定による Web 通信構造分析, 信学技報 NS2013-84.
[17] J. Ott, M. Sanchez, J. Rula, F. Bustamante, Content Delivery and the Natural Evolution of DNS, ACM IMC 2012.
[18] PlanetLab, https://www.planet-lab.org/
[19] Quantcast, http://www.quantcast.com/top-sites-1.
[20] J. Ravi, Z. Yu, and W. Shi, A survey on dynamic Web content generation and delivery techniques, Elsevier J. Network and Computer Applications, 32(5), pp.943-960, 2009.
[21] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann, The new web: characterizing AJAX traffic, ACM PAM 2008.
[22] A. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante, Drafting Behind Akamai: Inferring Network Conditions Based on CDN Redirections, ACM Trans. Networking, 17(6), pp.1752-1765, 2009.