東海大学紀要産業工学部２（2009年）73頁～80頁
Bull. School of Industrial Engineering Tokai Univ., 2 (2009) pp.73 -80

クローラの影響を考慮した学科作成 Web サイトのアクセス解析

松本 欣也*

Access Analysis of Department-Making Web Site with Considering Crawler Influence

by

Kinya MATSUMOTO

(Received: 19 OCTOBER 2009, Accepted: 22 FEBRUARY 2010)

Abstract

The purpose of this study was wide range analysis of access logs which were generated by department-making Web site. We analyzed the access log files including the activity of 19 months. In that case, crawler access was distinguished by methods of the clear definition. According to the analysis, the number of total access was 261,058 which include crawler access of 43.1%. This paper treats both of crawler and non-crawler analysis of the number of total access, distribution of access time, time transition of the number of unique IP address, mean number of access, source of access, kinds of top level domain (TLD), and estimation of main access region by using provider's domain name.

Key Words: Access analysis, Web, Crawler, Department-making site

１．はじめに

1.1　研究の背景

電子知能システム工学科では、東海大学公式ホームページとは異なる学科作成ホームページサイト[1)]を運用している。本サイトは、教育現場の日々の様子を学外に紹介するという公式ホームページに対する補完的な情報提供の役割を持っているが、公式ホームページ経由では得られない学科によるアクセス動向からの情報収集の役割も併せ持つ。

学科作成ホームページサイトのアクセスログの解析方法を明確にし、どのような項目が抽出できるのかを具体的に示すことは、研究教育機関からの情報発信を正確かつ効果的にする資料としての学術的な価値をもつ。このような研究例は数件の報告[2),3)]があるが、近年登場してきたクローラ（crawler）と呼ばれるWebサイトを自動巡回するソフトウェアロボットによるアクセスの影響を詳しく述べた文献がない。クローラが学科作成サイトのアクセス分析にどのような影響を与えているのか、実データを踏まえた分析が必要とされている。

本紀要では、学科作成ホームページサイトで取得した2008年2月26日から2009年8月31日までの19ヶ月間のアクセスログの解析を行うが、特にクローラからのアクセスの分離や、人間およびクローラのアクセスの傾向について実データに基づいた報告を行う。解析に用いたツールは、ブラックボックス化を防ぐため、特定の解析パッケージを使用せず UNIX の基本コマンドを用いて手法を明示する。

クローラからのアクセスの傾向を知り、人間からのアクセスを分離した解析を行うことで、①教育機関である学科の広報面に対してどのような要望があるか、②アクセス時期やアクセスされる内容は何かを明らかにでき、教育研究機関として正確で効果的な Web サイトの構築が期待できる。

２．解析方法

2.1 学科ホームページの構成

ホームページの管理を効率化するため、電子知能システム工学科ではHTMLベースのホームページからCMS（content management system）であるXOOPS[4)]ベースのホームページへ切り替えを行った。この学科作成 Web サイトは2008年2月26日に運用を開始した。アクセス解析に先立って、サイトの構成について述べる。まず、画面構成をFig. 1に、内容の構成をTable 1に示す。

* 産業工学部電子知能システム工学科准教授

Fig. 1 Snapshot of department-making Web site

XOOPSベースのWebサイトはブロック単位で情報と機能を集約する。Fig. 1の左端がナビゲーションの機能を果たし、中央が主たるコンテンツ、右端が入学希望者向け情報とナビゲーションになっている。アクセス解析を行うことにより、どのブロックへのアクセスがどの時期に多いか等の分析が期待される。また、Fig. 1中央の学科最新ニュースのブロックでは、各ニュース記事のアクセス回数が記録されており、どのようなニュースが期待されているかを調べることができる。

Table 1 Menu of department-making Web site on Aug. 2009

トップ	
├ メインメニュー	┐
├ インターネット大学	├ 左端
├ カレンダー	┘
├ 学科の最新ニュース	┐
├ 在学生向け情報	├ 中央
├ 産業界の最新ニュース	┘
├ 学科紹介アニメ	┐ 右端
└ オープンキャンパス情報	┘

2.1 アクセスログ

アクセス解析の手法は解析元となるデータの取得方法に応じて分類でき、アクセスログ型、パケットキャプチャ型およびビーコン型が知られている。後述の２方式は、Webサイトの開設時からパケットキャプチャを行ったり、Webコンテンツに JavaScript 等を埋め込んでソフトウェア的なビーコン発生源を挿入する等の特別な準備が必要である。近年公開された Google Analytics[5]は、ビーコン型に分類される。これに対し、アクセスログ型の解析ツールは JavaScript や Cookie を使用しないため、個々のセッションの管理を正確には行えない欠点があるが、解析手法が単純で、解析手法を変えながら何度も解析が可能である利点をもつ。本紀要では、アクセスログ型の解析を行った。用いたログファイルの記録項目を Table 2 に示す。

Table 2 A sample of record in access log file

項目名	値
接続元IP	216.129.119.40
ログイン名	-
ユーザ名	-
時刻	[01/Aug/2009:00:03:42 +0900]
メソッド	"GET /modules/… HTTP/1.0"
状態コード	200
送信バイト数	122296
リファラー	-
ユーザエージェント	"Mozilla/5.… /robot.html)"

アクセス解析には２つの目的、すなわちサーバパフォーマンスの計測とマーティングのためのデータ取得がある。本学科のサイトはアクセス数が多くないため、本報告ではサーバパフォーマンスの計測は実施せず、クローラの影響を考慮した総アクセス数の推移、閲覧に来たコンピュータ数（独立なＩＰアドレス）の推移、閲覧の多い時間帯、閲覧者の地域、および要求サービスの時間変化を抽出することに重点を置いた。

アクセス数については、アクセスログの行数を計数することで総アクセス数を得ることができ、Table 2 の時刻で分析を行うことで利用頻度の高い時間帯が得られる。

2.2 ドメイン分類

Table 2 の接続元 IP から IP アドレスを抽出し、DNS の逆引きによりドメイン名が得られれば、ドメイン名に含まれる文字列からアクセス元の情報がわかる。さらに、ドメイン名の最後尾にある TLD(top level domain)のうち ccTLD(country code TLD)を用いれば、国を識別できる。また、日本のドメイン(jp)は、属性型（組織種別型）、地域型および汎用のドメイン名から成っており、組織や地域に関する情報が抽出できる。ただし、DNS の変換テーブルは動的に更新されているため、過去の履歴である IP アドレスを全て変換することはできない。また、最初から DNS に未登録の場合もあるため、アクセスログ中の全IPアドレスにドメイン分類を行うことはできない。

2.3 サービス分類

利用者が閲覧を希望するサービスの情報は Table 2 のメソッド項目から得られる。どの時期に、どのようなサービスの閲覧が多かったか等のサービス分類が可能である。その際、全ての閲覧者がユーザ登録者であれば利用者の年齢や性別が自己申告ではあるものの確定でき、詳細なサービス利用の追跡が可能となる。しかし、最初の閲覧の際に個人情報の入力が必須になること、クッキーの利用が前提になること、および個人情報の管理業務が発生することから、学科作成ホームページサイトでは

ユーザ登録による閲覧を実施していない。よって、Table 2 のログイン名とユーザ名は使用しない。

３．解析結果

3.1 アクセス数

使用したログファイル httpd-access.log.*（ここで*は 0～18） には、Web 閲覧プログラムからのページ要求が記録されており、閲覧者（クローラを含む）のアクセス要求の履歴になっている。本紀要では、ログファイルの行数を全アクセス数と定義し、その推移を最初に調査した（Fig. 2 の total）。

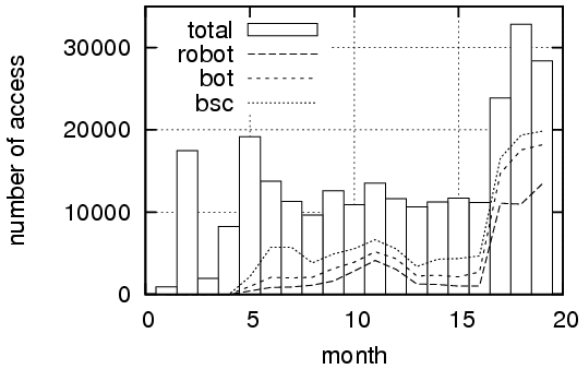

Fig. 2 Number of access on the Web site

全アクセス数にはクローラからのアクセスが含まれるほか、１ページの表示で多数のアクセスが記録される場合もあり、閲覧したページ数とは一致しない。利用者数や閲覧したコンピュータの台数に関する議論は 3.3 節で行う。ログファイルが１月毎に作成されるため、Fig. 2 は月単位でのアクセス数の推移を示しており、2009 年 8 月末（month=19）までの累積アクセス回数は 261,058 件（Fig. 2 total の合計）であった。

クローラは、robots、bots、spiders および harvesters とも呼ばれ、検索サイトに登録するためのデータを取得するプログラムである。これらからのアクセスは情報を広範囲に広める目的からは好ましいと考えることもできるが、人間の閲覧行動を前提としたアクセス分析を行う場合には除外して分析する必要がある。

クローラからの標準的なアクセスは robots.txt ファイルを GET メソッドによって取得するもので、ログファイル（Table 2）のメソッドとユーザエージェント項目を用いれば識別できる。しかし、クローラからの標準的でないアクセス、すなわち知的な振る舞いをするクローラからのファイルの要求については要求内容が未知であり、クローラがブラウザ情報に痕跡を残さない限り、人間からのアクセスと区別することは困難である。これは Web アクセス解析分野における近年の研究テーマであり、本紀要の範囲を超える。

本紀要では、クローラアクセスの大部分を分離し、人間からのアクセスがほとんどを占めるログファイルを生成して、これに対する解析を扱う。本紀要では、クローラアクセスを分離する方法として文字列の一致による方法を用いた。この手法は、方法（M-1）に示すように grep 指令を用いてログファイルの各行に含まれる文字列を探索する方法であり、文字列に”robot”を用いたもの（Fig. 2 の robot）、”bot”を用いたもの（Fig. 2 の bot）、および”bot”、”spider”、”crawl” のいずれかを含んだもの（Fig. 2 の bsc）を用いた。その際、大／小文字の区別はせず（－i オプション）、学科作成サイトのコンテンツにこれらの文字列は使用していない。本手法によるクローラの分離度については 3.6 節で述べる。なお、(M-1) の’¥’ 文字は継続行を表している。

```
grep －h －i 文字列　httpd-access.log.* |　¥
        tee robot-access.log | wc –l                (M-1)
```

方法 (M-1) によるクローラからの総アクセス件数は、robot が 54,961 件、bot が 83,595 件、bsc が 112,560 件であった。bsc の該当件数が最多であることから以後のクローラアクセスの分離には bsc を用いた。

クローラアクセスは 2009 年 8 月期（month=19）で 69.9%（19,831/ 28,364）、総アクセス数に対しては 43.1%（112,560/ 261,058）に達している。この数値は、人間の閲覧を前提としたアクセス分解析を行う場合には無視できない量であり、ログファイル型のアクセス解析を行う際にクローラアクセスの分離が必要なことを明確に示すデータといえる。

クローラアクセスの傾向を Fig. 2 から見ると、サイト開設後４ヶ月付近から順次増加している。学科からは検索サイトへ掲載の申し出を行っていないので、ネットワーク上に突然発生したサイトをクローラが発見していく過程を観察する１つのデータといえる。ここで、Fig. 2 において開設当初の４ヶ月はメンテナンスのために停止していた期間が含まれること、毎年 8 月に数日間ネットワーク接続が切断されたことを付記する。

3.2 アクセス時間帯

利用者が学科作成 Web サイトを閲覧する時間帯は、クローラからのアクセスとは異なると予想される。grep、awk、sed および wc 指令を用いて、アクセスログの日時欄を集計し、時間帯ごとの頻度分布を作成した (Fig. 3)。その際、クローラアクセスの抽出には前節 3.1 に示した bsc を用い、Fig.3 以降はこの方法により分離したデータを crawler と表記する。クローラアクセス（crawler）と非クローラアクセス（non crawler）はアクセス時間帯の分布が明確に異なることが Fig. 3 から分かる。すなわち、クローラアクセスは時間帯を問わず4,000～7,000 件／時

間のほぼ一様な分布を示すが、非クローラアクセスは15 時（日本標準時）と 23 時にピークをもち午前 5～6 時台で最小となる特徴的な分布を示す。この分布は、昼間に活動する人の生活サイクルと合致しており、bsc を用いた方法（M-1）によるクローラアクセス分離手法の正当性を示唆している。

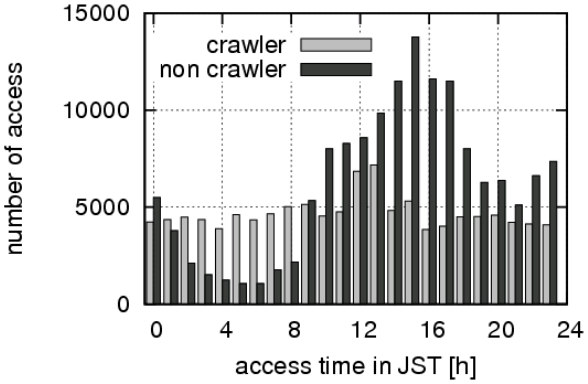

Fig. 3 Time-zone analysis of Web access

3.3 ユニーク IP アドレス

個人でも複数のコンピュータでブラウザを利用することから、コンピュータ毎に付される独立な IP アドレス（ユニークＩＰアドレス）の数は利用者数を直接に表すものではない。しかし、クッキーや認証を用いない Web サイトにおいては、利用者数を推し量る数値として有用である。

そこで、各相対月ごとに過去ログを全て積算し、その中から抽出したユニーク IP アドレス数の変化を探った（Fig. 4 の total）。また、１ケ月間のログのみから抽出したユニーク IP アドレス数も抽出した（Fig.4 の month）。Fig. 4 の crawler（破線）は、クローラのユニークＩＰアドレス数を示す。Fig. 4 の total、crawler および month の変化の様子は、いずれも Web サイトの注目度を測る指標の１つと考えられる。

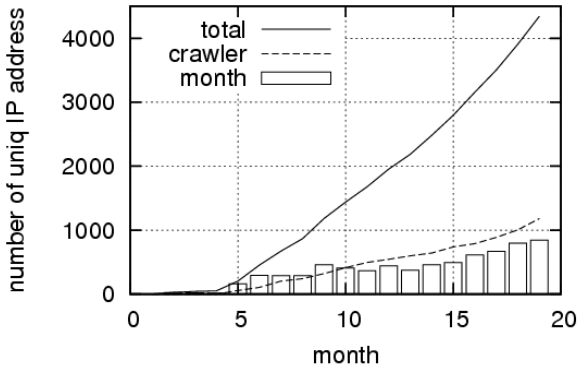

Fig. 4 Number of unique IP address

Fig. 4 の total から、学科作成サイトを閲覧しているコンピュータのユニークＩＰアドレスは、開設後４ヶ月以降からほぼ直線状に増加し、最終的に 4,341 件に達した（2009 年 8 月 31 日現在、month=19）。直線状の傾向は毎月ほぼ一定の割合で新規のコンピュータからのアクセスが発生していることを示す。ただし、プロバイダを利用した閲覧の際に異なるＩＰアドレスが割り振られることがあり、直接に利用者数の増加を示してはいない。

クローラのＩＰアドレス数（Fig. 4 の crawler）は開設後４ヶ月付近から増加しはじめ、total よりもゆるやかな増加傾向を示した。これは、クローラは 1 台で多数のアクセスを発生するためと考えられる。本解析により、現在明確にクローラとわかるユニークＩＰアドレスは 1,181 件でユニークＩＰアドレス全体の 27.2%である。

Fig. 4 から月毎のユニークＩＰアドレス数が得られたので、ＩＰアドレスあたりに発生したアクセス件数に注目し、式(1)により月別平均アクセス数 η の推移を調査した（Fig. 5）。横軸は開設後の経過月を示す。

$$\eta = \frac{(\text{該当月のログ件数})}{(\text{該当月のユニークＩＰアドレス数})} \qquad (1)$$

η は、閲覧者が開いたページ数と相関があり、Web サイトの注目度を示す１つの指標と考えられる。Fig. 5 の total は、クローラを区別せずに式(1)から得た η である。一方、crawler はクローラのみを抽出した月別ログファイルから得た平均アクセス数 η_c、non crawler は非クローラアクセスのみを抽出した月別ログファイルから得た η_n を表す。

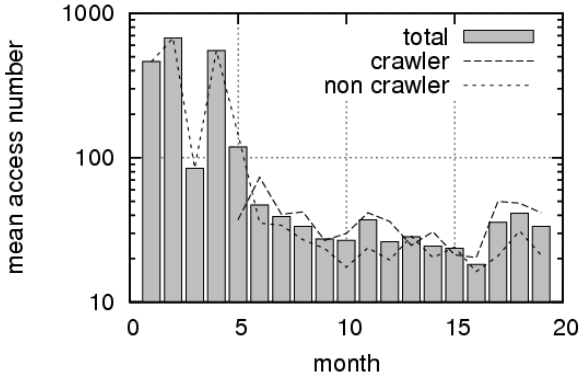

Fig. 5 Mean access number by single IP address

開設直後の４ヶ月はメンテナンスのために、少数のＩＰアドレスから多数のアクセスがあったことが分かる。それ以降は、予想と異なり crawler も non crawler も明らかな差異を示さず η =20～70[件／IP]となった。細かく見れば、直近の３ヶ月間は η_c が η_n の 2 倍の平均アクセス回数となっている。

Fig. 5 から、開設直後の注目度が徐々に減少し、2009 年 5 月（month=16）に底に達し、6～7 月（month=17～18）にかけて回復している様子が読み取れる。この変化は、Web サイトの運用の状況と合致する。今年 3 月

（month=14）に平均アクセス回数の減少を察知したため停滞していた記事更新を4月（month=15）以降に再開した経緯がある。この傾向はFig. 4からは読み取れず、Webサイトへの注目度を示す指標としては平均アクセス回数の方が相応しいという結果を得た。

3.4 クローラアクセスのドメイン分析

どのような利用者が閲覧しているのかをドメイン名から探る準備として、アクセス数の多いＩＰアドレスを調査した[方法（M-2）、Table 3]。方法（M-2）は方法（M-1）で分離したクローラアクセスのログrobot-access.log.*に対してＩＰアドレス毎のアクセス数を集計して多いものから表示する。

```
awk '{print $1}' robot-access.log.* | ¥
sort | uniq –c | sort –r –n                    (M-2)
```

Table 3 Top 10 accesses from crawler IP

IPアドレス	アクセス数
119.63.194.32	12,435
38.99.44.101	9,851
216.129.119.40	9,167
119.63.193.103	4,717
119.63.193.75	4,394
61.247.222.55	3,181
61.247.222.45	3,165
61.247.222.56	3,144
61.247.222.54	3,054
61.247.222.44	2,749

Table 3には似通ったＩＰアドレスが多く見られる。これは、クローラが複数のコンピュータを用いてクローリングしていることを示唆するが、ＩＰアドレスからはグループの判別ができない。そこで、DNSの逆引きからドメイン名を取得してグループ化を試みた。取得できたドメイン名は803件で、取得率は68.0%であった。これらのドメイン名に対して組織ドメインを用いたグループ化を行った。主なグループをTable 4に示す。

Table 4 Main access from crawler domain

ドメイン名	N	アクセス数	比率
crawl-*.naver.jp	23	26,639	23.7
crawl-*.cuil.com	4	19,037	16.9
*.search.msn.com	285	13,293	11.8
*googlebot.com	81	8,548	7.6
*.crawl.yahoo.net	291	1,041	0.9
取得不可能	379	33,040	29.3

NはユニークなＩＰアドレス数、比率はクローラからの総アクセス数に対する百分率

Table 4から、同一の組織ドメインから4～291台のコンピュータを用いてクローラアクセスが行われていたという実態が明らかとなった（Table 4）。ここで、１つのＩＰアドレスが１台のコンピュータを表すと仮定している。

Table 4のコンピュータ（1,063台）からのアクセス数を集計すると101,598件であり、クローラによる全アクセス（3.1節より112,560件）の90.2%に達する。このことから、方法（M-1）で分類したクローラアクセスの発生源はTable 4でほぼ網羅したといえる。

3.5 非クローラアクセスのドメイン分析

全アクセスログから方法（M-1）によりクローラアクセス（3.1節より43.15%）を分離したが、残りの56.9%についてドメイン名を取得して分類を行った。これらのアクセスに含まれていたユニークＩＰアドレスは3,487件で、このうちドメイン名が取得できたものは2,815件（80.7%）であった。ドメイン名の取得率はクローラアクセスに比べ12.7高かった。

取得したドメイン名のTLDを分類したところ54種類が確認された（Table 5）。Table 5に含まれるccTLDは45種類あり、小／大文字表示が同一と判断すれば43カ国となる。同一国家で10以上のドメイン名を記録した国は、日本(jp)、ブラジル(br)、ドイツ（de）、ロシア連邦（ru）、および中国（cn）である。

その他のTLDについては、gTLD（generic TLD）が5種類（小／大文字表記が同一とすれば4種類）、Infrastructure TLDが1種類、Reserved TLDが1種類、不明が2種類（pacificswell、adsl）であった。

TLDのうち、多くのドメインを持つものは、jp (1428)、net(899)、com(399)であった。利用状況の追跡を行う際は、これらの分析を行えばよい。

Table 5 Detected TLD with number of domain

```
jp(1428) net(899) com(300) br(40) de(18)
ru(11) cn(10) it(9) ca(8) tw(7) pl(7) sg(5)
uk(4) nz(4) us(3) pt(3) pk(3) mx(3) info(3)
in(3) eg(3) au(3) tr(2) th(2) se(2) org(2)
kr(2) fr(2) ee(2) cz(2) COM(2) ws(1) ua(1)
sk(1) rs(1) pacificswell(1) nl(1) md(1)
ma(1) localhost(1) jo(1) ie(1) hu(1) gr(1)
fi(1) dk(1) cy(1) be(1) at(1) arpa(1) ar(1)
adsl(1) UA(1) CA(1)
```

Table 6に、jpドメインのサブドメインについて分類した結果を示す。jpドメインには34種類のサブドメインが検出され、ドメイン数が多い順に、ne.jp、or.jp、ac.jp、bbiq.jpおよびco.jpであった。BBIQは九州電力グループのQTNet（本社 福岡市）が提供する光ブロードバンドインターネット接続であり、BBIQドメインが上位にあることは九州内からのアクセスが多いことを示唆

する。第３レベルのサブドメインまで調査を行うと、ドメイン数の多いものから ocn.ne.jp(245)、enjoy.ne.jp (213)、plala.or.jp(102)、ppp.bbiq.jp(83)、mesh.ad.jp (63) 、infoweb.ne.jp(58)他の順であった。この結果は、非クローラアクセスにおいて閲覧者が使用したプロバイダを反映した結果と考えられる。

Table 6 Detail of jp domain

ne.jp(887)	or.jp(118)	ac.jp(112)
bbiq.jp(85)	ad.jp(75)	co.jp(72)
bbexcite.jp(18)	ed.jp(13)	j-cnet.jp(7)
go.jp(6)	m-zone.jp(5)	commufa.jp(4)
zoot.jp(2)	riken.jp(2)	naist.jp(2)
kumamoto.jp(2)	wako-e.jp(1)	w-lan.jp(1)
techno-arc-shimane.jp(1)		syswave.jp(1)
starflyer.jp(1)	srg.jp(1)	shutoko.jp(1)
lolipop.jp(1)		lg.jp(1)
jeevessolutions.jp(1)		jaxa.jp(1)
iwate.jp(1)	itscom.jp(1)	hiroshima.jp(1)
dsn.jp(1)	doubleroute.jp(1)	cv-net.jp(1)
chiba-u.jp(1)		

Table 6 の ne.jp には OCN プロバイダの ocn.ne.jp が 200 件含まれているが、これらのドメイン名には kumamoto.ocn.ne.jp のように地域名が含まれている。この特徴を利用すると、アクセスログからは知ることができない閲覧者の地域分布を推定できると考えた。この考えに基づき、ocn.ne.jp ドメインに含まれる文字列から地域を分類した結果を Fig. 6 に示す。

Fig. 6 の結果が全アクセスに対する結果と同一になるためには、ＯＣＮ経由のアクセスが占める割合が全国各地で均等でなければならない。このデータは未調査であるため、Fig. 6 は全アクセスに対する非常に荒い推定値を与える。

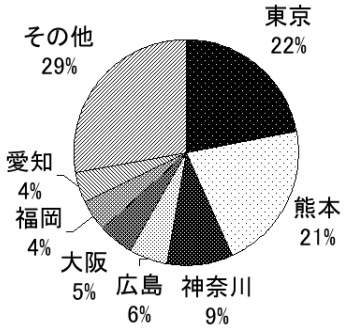

Fig. 6 Access region analyzed from OCN domain

Fig. 6 によれば、東京・神奈川からのアクセスが 32.3% 熊本・福岡からのアクセスが 25.7%で過半数のアクセスを占めた。その他の内訳は、括弧内に独立なドメイン数を書き、北海道(8)、埼玉・宮城(7)、秋田(5)、栃木(4)、静岡・京都・香川・石川・茨城(3)、鹿児島・兵庫・千葉(2)、山梨・山口・富山・沖縄・岡山・長崎・長野・宮城・岩手・群馬(1)であった。以上から、九州圏内において熊本・福岡以外からのアクセスが少ないことが明らかで、これらの地域でも OCN の利用者が他県と同様にいるならば学科作成サイトの認知度が低いといえる。

3.6 要求サービスの分析

Table 2 のメソッド項目には閲覧者が希望した情報源のＵＲＬが記載されるため、要求サービスの分析が可能である。

まず、クローラアクセスの要求サービスを分析した。クローラの基本的な動作は”GET /robots.txt HTTP/1.1”および”GET /robots.txt HTTP/1.0”である。これらを抽出すると、各々8,925 件および 1,514 件であった。これは GET メソッドの 9.3%（10, 439/ 112,559）にしかならず、基本的なクローラアクセスが少ない結果となった。

クローラアクセスにおける他の GET メソッドは 102,120 件あり、利用頻度の高いサービスを抽出すると ”GET /modules/bulletin”(51,451)、”GET /modules/piCal”(17,590) 、”GET /modules/weblogD3” (6,388) 等であった。これらのメソッドに含まれる/modules 等の文字列は XOOPS に特有のディレクトリであり、ホームページに含まれる各ブロックの機能を実現するスクリプトを格納している。クローラアクセスにおいて、基本的なアクセスよりこれらの機能への要求が圧倒的に多いことは、知的なクローラアクセスが主流になっていることを示している。

次に、非クローラアクセスの要求サービスについて調査を行った。GET メソッドは 140,882 件あり、非クローラアクセス全体の 94.9%であった。利用頻度の高いサービスは”GET /themes/sangyokou”(24,852)、”GET /modules/piCal”(18,909)、”GET /uploads/fckeditor”(10,271)、”GET /modules/pico_labo”(8, 899) 等であった。

非クローラアクセスの基本的な要求サービスはトップページの表示である。トップページの要求があると”GET / HTTP/1.1”または”GET / HTTP/1.0”のメソッドが記録されるので、方法（M-3a）で抽出を行った。

grep "GET / HTTP/" norobot.log.* | wc –l　　(M-3a)

結果は 7,769 件(非クローラアクセス全体の 5.2%)であった。ここで、非クローラアクセスの閲覧者はインターネットエクスプローラ（MSIE）か Firefox ブラウザを用いて閲覧していると考えられるので、方法（M-3b）により MSIE と Firefox ユーザに限定して抽出し、4,362 件を得た。

grep "GET / HTTP/" norobot.log.* | grep ¥
–e "MSIE" -e "Firefox" | wc –l　　(M-3b)

アクセスログを観察すると、この他に学科作成サイトのサーバ内部からのアクセスが Apache からのアクセスとして 3,000 件記録されていた。以上から、ブラウザとサーバ内部アクセスの合計が、非クローラアクセスのト

ップページ要求の 94.8%（7,362/7,769）という高い割合になり、クローラからのアクセスが混在する余地がほとんどない。つまり、非クローラアクセス中に分離しきれずに残ったクローラアクセスが僅かであることを示している。

以上の分析結果を踏まえて、非クローラアクセスの全データを使って要求サービスの分析を行った（Table 7）。その際、要求サービスを識別するためにメソッドに含まれるサービス固有の文字列を知る必要があったので、トップページの全リンクをクリックしてログに記録させる実験を行った。

Table 7 Analysis of service demanded on the Web site

トップ	アクセス数
├ メインメニュー	―
│ ├ ホーム	654
│ ├ 学科紹介	607
│ ├ 研究室紹介	1,108
│ ├ ニュース	453
│ ├ プログ	339
│ ├ カレンダー	180
│ ├ ヘッドライン	213
│ ├ コンタクト	263
│ └ リンク集	202
├ インターネット大学	―*
├ カレンダー	53
├ 学科の最新ニュース	3,850
├ 在学生向け情報	―
│ ├ 写真 51	14
│ └ 写真 43	160
├ 産業界の最新ニュース	―*
├ 学科紹介アニメ	―*
└ オープンキャンパス情報	―
├ 学部ムービー	―*
├ 入学広報	―*
├ パンフ	―*
├ 写真（学科）	64
├ 写真（研究室１）	54
├ 写真（研究室２）	39
└ 写真（研究室３）	41

―*印は外部リンクのためアクセス履歴がないもの
―はアクセス数を管理しないもの（トップページの表示で情報が読めるため）

Table 7 から、全ての要求サービスの中で学科の最新ニュースの閲覧が多く、次に研究室紹介、学科紹介の順であった。ただし、外部へのリンクについてはアクセス記録がなく不明であった（Table 7 *印）。主な要求サービスが明らかになったので、これらのサービスの月別利用情報を抽出した（Fig. 7）。Fig. 7 に、学科の最新ニュース（news）、学科案内（intro）および研究室案内（labo）の月別アクセス状況を示す。

利用サービスのうち、メニュー項目（intro, labo）はコンテンツ項目（news）よりアクセスが少なく同じような月別アクセスの傾向を示している。コンテンツ項目の news は閲覧要求のオーダが異なるとともに閲覧要求が激しく変動している。これは、ニュースのアップデートの影響と考えられる。

以上の結果から、現状のページ構成のままで学科作成サイトの利用者数を効果的に増大させるには、学科の最新ニュースの充実が効果的な方法である。

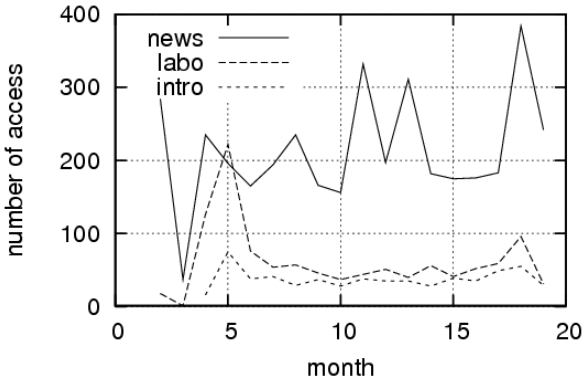

Fig. 7 Access number of main services

3.7 訪問経路

検索サイトから到着した閲覧者のアクセスログには、検索サイトで入力した日本語文字列が含まれている。この特徴を利用すれば、閲覧者がどのようなキーワードで学科サイトに達したか、すなわち学科をどのように捉えているかというデータが得られる。非クローラアクセスの中で検索サイトから訪問した閲覧者のアクセスを抽出するために方法（M-4）を用い 201 件を抽出した。ログファイル中の文字列はパーセントエンコードされているので、デコードした後に文字列の頻度を分析した。

```
grep "GET / HTTP/" norobot.log.* | ¥
grep –e "MSIE" –e "Firefox" | ¥
grep "search?"    > res-search.txt          (M-4)
```

検索サイトに入力した文字列は複数であることが多かった。使用頻度の高い文字列は、東海大(133)、電子知能(68)、産業工(38)、インターネット大(23)、熊本(21)、CG(9)、中嶋(9)、研究室(４)、知能システム工学(4)、でかちゃれ(3)、熊本キャンパス展(2)、情報化フェア(2)、ET ロボコン(2)、工学部と産業理工学部の違い(1)、パンフレット プレビュー(1)、アニメーション 熊本 大学(1)であった。以上から、これまでの閲覧者に関しては、検索サイトを利用する場合はあらかじめ東海大で絞り込んでおり、漠然と学問のキーワードを並べて辿り着いたのではないことが明らかとなった。この結果は、大学内の学部学科の系統を明示した中で、学科の個性を伝えるページ作りが有効であることを示している。学術的な内容を中心とした場合は、現在の検索の様子からは検索に

かからないといえる。

次に、東海大学公式サイトから来訪している利用者についての分析を行った。公式サイトからの来訪者は方法（M-4）の grep “search?” の部分を方法（M-5）に変えて実行することで得られ、1,147 件を抽出した。これらの閲覧者について月別の頻度分布を求めた（Fig. 8）。

grep "_and_intellige/" (M-5)

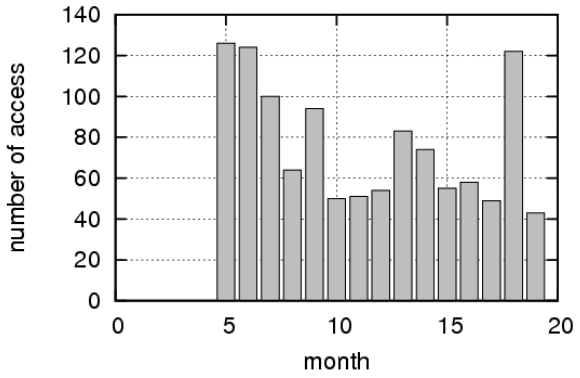

Fig. 8 In-coming access from formal University Web

公式サイトからの訪問者は2008 年 6 月（month=5）から記録されており、120 件／月を越えたのは 2008 年 6 月、7 月（month=6）および2009 年 7 月（month=18）であった。これらの時期には、学部学科に関する公式ホームページの情報を閲覧しつつ、より詳細な情報を得たいために学科作成 Web サイトに来訪する閲覧者が増加している。

４．まとめ

本紀要では、電子知能システム工学科の学科作成ホームページサイトのアクセス解析を、クローラからの影響を考慮して行った。解析手法にはアクセスログ型の手法を用いた。

アクセスの総数と月別の状況についてクローラおよび非クローラのアクセスに分離して分析を行い、2009 年 8 月 31 日までの累積アクセスが 261,058 件（うちクローラアクセス 43.1%を含む）であることを明らかにした（3.1 節）。クローラの分離に関しては、その手法を明示した。また、アクセスの時間帯を分析し、クローラと非クローラのアクセスの傾向に明確な違いがあることを明らかにした（3.2 節、Fig. 3）。

何種類の IP アドレスからアクセスがあったかをユニーク IP アドレスにより分析した（3.3 節）。ユニーク IP アドレスは学科作成サイトの顧客であり単調な増加を示した。ユニーク IP アドレス数から求めた平均アクセス数には、コンテンツの変更を反映した変化が検出され Web サイトの注目度を得る有用な指標と考えられる。

アクセスの発生源について、クローラアクセス（3.4 節）と非クローラアクセス（3.5 節）についての分析を行った。その結果、5 つの組織ドメインを主要なクローラアクセス発生源として特定した。全アクセスの 56.9% を占める非クローラアクセスについて、TLD を分析し 43 ケ国からのアクセスを検出し、国別のドメイン数を得た。jp ドメインのサブドメインを集計し、プロバイダのドメイン名がもつ特徴を利用してアクセス地域の推定を行い、東京、熊本および神奈川からのアクセスが過半数を占める推定結果を得た。

要求サービスの分析からは、クローラによる知的なアクセスが多数を占める様子を明らかにした（3.6 節）。非クローラアクセスの要求サービス分析からは、主たる要求サービス（Table 7）が明らかとなり、月別アクセス状況が抽出できた（Fig. 7）。

最後に、検索サイトから訪問した利用者の履歴を使って、学科作成サイトを検索した文字列を明らかにした。加えて、大学の公式サイトから訪問した利用者の月別アクセス状況を抽出した（Fig. 8）。

19 ヶ月に渡って収集された学科作成サイトの履歴から以上の具体的な数値や傾向が明らかになった。本紀要の分析により、利用者の要求サービスや閲覧の多い時期などの知見に基づく学科作成サイトの運営が可能になった。

謝辞　本論文は、電子知能システム工学科ホームページのアクセス履歴を使用させていただきました。運用にご協力いただいている学科教職員の皆様に感謝いたします。特に、同ホームページの構築とサーバ管理を担当されている中嶋卓雄先生、および学科主任井手口健先生に感謝します。

引用文献

1) http://www-id.ktokai-u.ac.jp/ (2009.9.14 現在)
2) 簗瀬洋一郎: 中京学院大学ホームページのアクセス解析,中京学院大紀要, 7(2),13,(2000), pp.31-37.
3) 保坂邦夫: 全入時代の広報戦略(6) 昭和女子大学の事例, 私学経営, 410, (2009), pp.44-51.
4) http://www.xoops.org/ (2009.9.14 現在)
5) Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber : “Bigtable: A Distributed Storage System for Structured Data”, proc. OSDI’ 06, (2006)., pp.205-218.