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

2009－ＤＰＳ－138 (30)
2009－CSEC－ 44 (30)
2009／3／6

Crawler の動作特性の統計的分析

服部祐一 中嶋卓雄
東海大学
〒862-8652 熊本県熊本市渡鹿 9-1-1
E-mail: 6aik1144@mail.tokai-u.jp, taku@ktmail.tokai-u.jp

概要

WEBサーバアプリケーションのアクセスログおよびWEBサーバに対するファイアーウォールのアクセスログから,ロボット型サーチエンジンに使われているCrawlerのアクセスログ部分を抽出する.そして,Crawlerのアクセス数とファイルの種類に対するCrawlerのアクセスの割合からCrawlerの動作特性について統計的な分析を行う.その結果から主要な4つのCrawlerの動作特性の違いについて検証する.

Statistical analysis of crawler's operation characteristic

Yuichi Hattori and Takuo Nakashima,
Tokai University
9-1-1, Toroku, Kumamoto, 862-8652, Japan
E-mail: 6aik1144@mail.tokai-u.jp, taku@ktmail.tokai-u.jp

We extract the access log of crawler used in robot type search engine from the access log of web server application and firewall's access log of web server's access. We conduct statistical analysis about an operation characteristic of crawler from the number of the access and a ratio of the access of crawler for the kind of the file. We inspect it about the difference of the operation characteristic of four main crawlers from the result.

1. はじめに

近年,GoogleやYahooなどのロボット型サーチエンジンで,WEBページの検索結果を上位に表示させることが重要視されており,企業などにとっては顧客獲得に関わることである.

本論文では,ロボット型サーチエンジンに使われているCrawlerのWEBページに対するアクセスについて着目し,WEBサーバアプリケーションのアクセスログとWEBサーバに対するファイアーウォールのアクセスログから代表的なCrawlerのアクセスを抽出する.その結果から各Crawlerにどのような動作特性があるかを統計的に分析する.

2. 代表的 Crawler

本論文では,下記の４つの代表的なCrawler

の動作特性について分析する.

- Googlebot

Googlebot とは,Google が使用している Crawler である[1].

- Yahoo! Slurp

Yahoo! Slurp とは Yahoo の Yahoo! Search Technology に使われている Crawler である[2].

- msnbot

msnbot とは MSN および Windows Live Search が使用する Crawler である[3].

- Baiduspider

Baiduspider とは,Baidu.jp の検索エンジンで利用されている Crawler である[4].

3. WEB サーバアプリケーションのアクセスログからの分析

3.1 使用したアクセスログ

今回使用したアクセスログは,2008 年 12 月 8 日から 2009 年 1 月 8 日までの WEB サーバアプリケーションである Apache HTTP Server の標準設定のアクセスログである.なお,このサーバの設定に関しては,ファイアーウォールの段階で中国と韓国のアクセスをブロックしている.

3.2 Crawler の抽出手法

今回行った WEB サーバアプリケーションのアクセスログからの Crawler の抽出手法は,WEB サーバアプリケーションのアクセスログから User Agent と IP アドレスの部分を抽出し,それを元に Crawler かどうか判別する.そして,Crawler であればそのアクセスログを抽出する[表 1].

表 1 抽出した UserAgent

Crawler	User Agent
Googlebot	Googlebot/2.1
	Googlebot-Image/1.0
	Googlebot-Mobile/2.1
Yahoo! Slurp	Yahoo! Slurp/3.0
	Yahoo! Slurp
msnbot	msnbot/1.1
	msnbot/1.0
	msnbot-media/1.1
	msnbot-media/1.0
Baiduspider	Baiduspider+
	BaiduImagespider

3.3 User Agent の確認

UserAgent は容易に偽装できるため,それらが偽装されてないか確認する.まず,抽出した IP アドレスに対し nslookup コマンドを用い,DNS サーバに問い合わせる. そして,IP アドレスに対するホスト名を取得し,偽装でないか確認する.これにより,Baiduspider 以外の Crawler に関しては,正規性を確認することができた. Baiduspider に関しては,ホスト名を取得することができなかった.
そのため,Baiduspider の IP アドレスに対して,whois コマンドを用いて,ドメイン情報を取得し,それを元に判別を行った.
これにより,Baiduspider の正規性も確認することができた.

3.4 Crawler ごとの日別統計

図 1 は,抽出した Crawler のアクセスログを日付別に集計したものである.その結果から,Goolglebot に関しては,他の Crawler に比べ,アクセスが少なく数値の変動を少ない.

Yahoo! Slurp に関しては,全体的には,BaiduSpider の次にアクセスが多いものが,この期間中に検索アルゴリズムのアップデートはあったもののインデックスのフルアップデートなどはなかったため特に目立っ

た変化は見られない[5].

msnbot に関しては,12 月 18 日あたりからアクセス数が若干増えている.その時期に msnbot の Crawler 自体のアップデート[7]があったが,アップデートされた Crawler の UserAgent である「msnbot/2.0b」が抽出した UserAgent にないためアップデートの影響ではないと考えられる.そのため,これは WEB ページの変更等によるものだと推測される.

Baiduspider に関しては,全体的にほかの Crawler と比べアクセスが多いものの,アクセス数にばらつきが見られる.

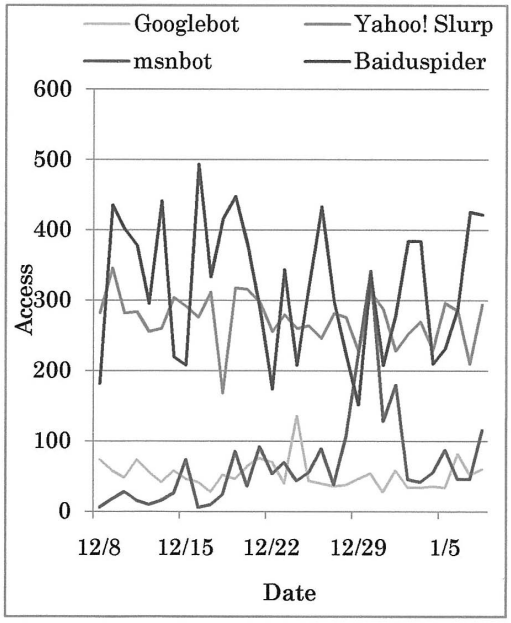

図 1 WEB サーバアプリケーションのアクセスログにおける Crawler ごとの日別統計

3.5 Crawler ごとのファイル種類別アクセスの割合

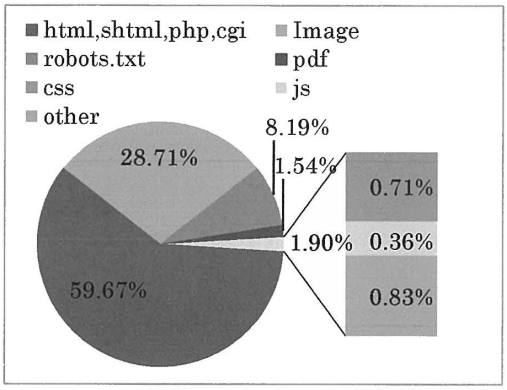

図 2 Googlebot におけるファイル種類別アクセス割合

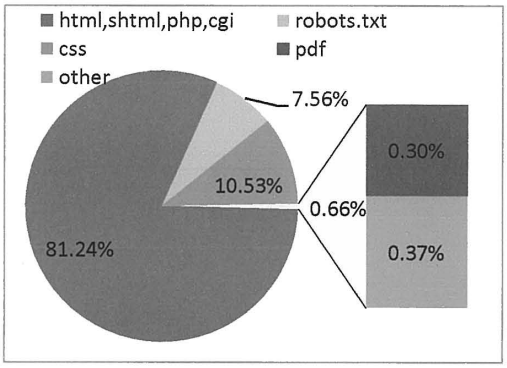

図 3 Yahoo! Slurp におけるファイル種類別アクセス割合

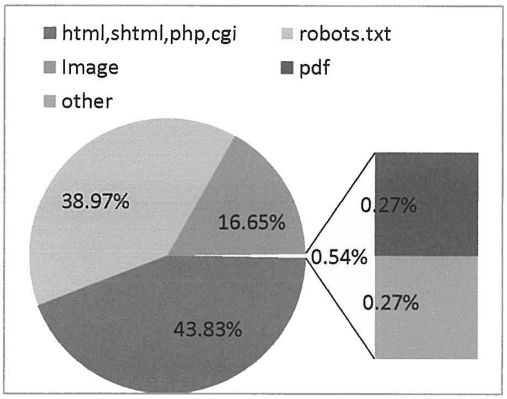

図 4 msnbot におけるファイル種類別アクセス割合

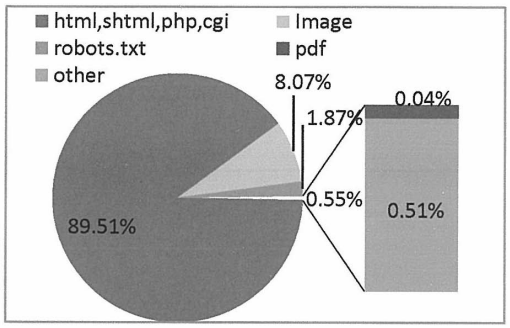

図 5 Baiduspider におけるファイル種類別アクセス割合

Crawler のアクセスについて,ファイル種類別という観点から,html,shtml,php,cgi を用いて記述された WEB ページを構築するファイル,gif,jpg,png,bmp などの画像ファイル,ロボット型サーチエンジンに対する命令を記述する robots.txt などに分けて分析した.なお,このサーバの WEB ページでは robots.txt による Crawler の回避などは行っていない.その結果,html,shtml,php,cgi などのファイルに対するアクセスは,Baiduspider が最も多く全体の 89.51%を占めている.その一方,msnbot は 43.83%しかなく,他の Crawler に比べ少ない.robots.txt に関しては,msnbot が 38.97%となっており,他の Crawler に比べ大幅に割合が大きい.robots.txt はそれほど更新が多いファイルではないため,msnbot の robots.txt に対するアクセスは多すぎると考えられる.

0.36%と僅かではあるが,Googlebot のみ通常の検索に用いるだけであれば必要がないと考えられる Javascript のファイルにアクセスしている.そのため,Googlebot は,Javascript のファイルもページの順位決めなどに利用している可能性がある.Yahoo! Slurp と Googlebot は CSS に対するアクセスがあり,特に Yahoo! slurp は 10.53%が CSS に対するアクセスの割合である.CSS は,通常の検索に用いるには必要のないものと考えられるのでこれも,ページの順位決めなどに利用している可能性がある.

画像ファイルに関しては,Yahoo! Slurp のみアクセスが全くない.Yahoo にもイメージ検索のサービスがあるため,画像ファイルに対するアクセスは必要なものと考えられるが,画像ファイルに対するアクセスがないことから推測すると,img タグの alt 属性の文字列,src 属性のファイル名などから検索を行っているのではないかと推測される.一方,他の Crawler では,画像ファイルに対するアクセスがあるため,画像ファイル自体も検索の対象となっているものと推測される.

3.6 アクセスログにおける動作特性

アクセスログ自体にも Crawler によっては,他と異なる動作特性が確認できた.Googlebot と msnbot の画像ファイルに対するアクセスは GET メソッドが用いられるが,Baiduspider は,HEAD メソッドを用いてアクセスしている.そのため,Baiduspider の画像に対するアクセスは,ヘッダ情報のみしか取得していないため Baidu.jp の画像検索には,ヘッダ情報が主に使われているものと推測される.

Yahoo! Slurp 以外の Crawler の HTTP プロトコルのヴァージョンは 1.1[9]なのに対し,Yahoo! Slurp のほぼすべてのアクセスの HTTP プロトコルのヴァージョンは 1.0[8]である.1.1 では一回の TCP/IP 接続につき,複数の HTTP リクエストとレスポンスができ,1.0 よりもサーバの負荷を減らすことができるなど効率的であるため,HTTP のヴァージョンは 1.1 を使うべきであると考えられる.

4. ファイアーウォールのアクセスログからの分析

4.1 使用したアクセスログ

今回使用したアクセスログは,2008年9月7日から2008年12月21日までのファイアーウォールのWEBサーバに対するアクセスログである.なお.サーバに関しては3で用いたサーバとは別のサーバである.

4.2 Crawlerの抽出手法

今回行ったファイアーウォールのアクセスログからのCrawlerの抽出手法は,WEBサーバアプリケーションのアクセスログの場合と異なり,IPアドレスを元に探索するしかないため, 3.3に用いたUser Agentの確認の手法を用いてホスト名とドメイン情報を元に抽出する.

4.3 Crawlerごとの日別統計

Goolebotに関しては,不定期に大幅にアクセスが増加しているが,どれも短期間でその後,元のアクセス数に落ち着いているので,Crawler自体のアップデートによるものとは考えにくい.

Yahoo! Slurpに関しては,9月末頃からアクセスが大幅に増加している.これは9月28日にYahoo! Search Technologyのインデックスのフルアップデートを開始した影響と考えられる[6].その後,10月15日頃にアクセス数が大幅に減少しているが,これも9月末のときと同様にインデックスのフルアップデートを開始した影響と考えられる.開始された日時は10月15日である[6].10月15日以降はそれほど大きな変動はみられないが,9月28日のアップデート以前に比べ,アクセス数は約2倍に増加している.

msnbotに関しては,不定期にアクセスが大幅に増加している期間があるが, 告知されているアップデートの日時とは,離れているためアップデートによるものではないと考えられる.[7]

Baiduspiderに関しては,他のCrawlerに見られるような集中的なアクセスの増加の傾向はみられない.

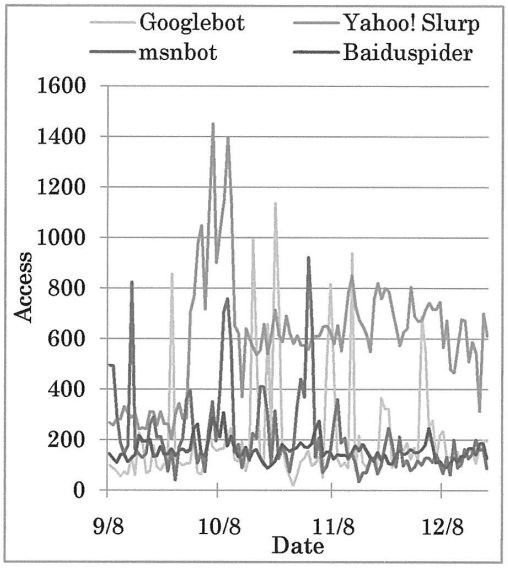

図 6 ファイアーウォールのアクセスログにおけるCrawlerごとの日別統計

5. まとめと今後の課題

本論文では,Crawlerの動作特性について,WEBサーバアプリケーションのアクセスログとファイアーウォールのWEBサーバに対するアクセスログから分析を行った.

今後の課題は,今回分析に使ったWEBサーバアプリケーションのアクセスログが約一ヵ月分でデータとして少なかったため,今後は,より多くの期間のアクセスログを用いて分析していくとともに,ファイルに対するアクセス間隔やサーバに対する負荷といった点からも分析していく.

参考文献

1. Verifying Googlebot - Webmasters/Site owners Help, http://www.google.com/support/webmasters/bin/answer.py?answer=80553&topic=15265, Google, January 2009
2. Yahoo's Web Crawler, http://help.yahoo.com/l/us/yahoo/search/webcrawler/, Yahoo!, January 2009
3. Windows Live ヘルプ http://help.live.com/help.aspx?mkt=ja-JP&project=wl_webmasters, Microsoft Corporation, January 2009
4. 百度ヘルプセンター, http://help.baidu.jp/system/05.html, Baidu Inc., January 2009
5. Yahoo!検索 スタッフブログ http://searchblog.yahoo.co.jp/, Yahoo Japan Corporation, January 2009
6. Yahoo! Search Blog, http://ysearchblog.com, Yahoo! Inc., January 2009
7. Live Search Webmaster Center Blog, http://blogs.msdn.com/webmaster/, Microsoft Corporation. January 2009
8. Hypertext Transfer Protocol – HTTP/1.0, http://www.w3.org/Protocols/rfc1945/rfc1945, W3C, May 1996
9. Hypertext Transfer Protocol – HTTP/1.1, ftp://ftp.isi.edu/in-notes/rfc2616.txt, W3C, June 1999