3T－1

WWWにおけるUser-Agent特定のための アクセスログ解析手法

河辺岳人†、宮崎正俊‡

†株式会社ＳＲＡ東北　‡岩手県立大学ソフトウェア情報学部

1．はじめに

World Wide Web (WWW) では、標準にのっとったシステムを用意することにより、情報提供者・情報利用者とも特定のベンダに依存することなくシステムの構築・利用が可能となっている。このため、旧来の情報システムと異なり、利用者の使用している端末ソフトウェアや取得後の情報の利用方法も多彩になってきている。

このようなシステムにおいては、情報提供者からは(1)利用者側の端末ソフトウェアに適した情報の提供、(2)情報自体、及びサーバーの意図しない利用からの保護、(3)情報に広告を追加する際の、その情報の露出度のより正確な測定、といった要求がある。

利用されている端末ソフトウェアの種類を正確に知ることができれば、こういった要求に答えることが可能である。

本稿では、端末ソフトウェアをより正確に特定するためのHTTPサーバーのアクセスログの解析手法について述べる。

2．アクセスログ解析手法

本稿では、端末ソフトウェアを特定・分類するため、図１に示すように２段階に分けてHTTPサーバーのアクセスログの解析を行う。以下に本稿にて用いる用語および解析手法と、解析にあたってどのようなデータが必要かを示す。

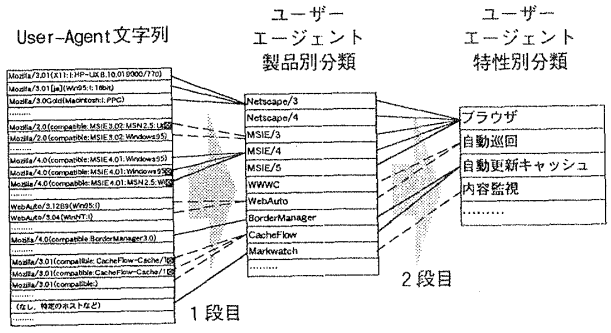

図１．ユーザーエージェントの２段階分類

Access Log Analysis Method for Distinguishing WWW User Agents
Taketo Kabe, Masatoshi Miyazaki
Iwate Prefectural University, Iwate, JAPAN

2.1 用語

ユーザーエージェント：本稿が対象としているWWWにおける端末ソフトウェア。ユーザーからの要求、もしくは自律的判断にもとづきHTTP[1]を用いてHTTPサーバーより情報を引き出す。

User-Agent文字列：ユーザーエージェントがHTTPにて情報を取得する際、通常はプロトコル中のUser-Agentヘッダに製品が識別できるような文字列を渡すことが求められている。本稿の目的であるユーザーエージェントの特定には基本的にこの情報を利用する。なお、通常のログ[3]ではこの情報は記録されないため、サーバーにて明示的に設定を行う必要がある。

ブラウザ：ユーザーエージェントのうち、ユーザーが直接操作してHTTPサーバーから情報を取得し、すぐに閲覧できる機能を持つもの。

検索ロボット[4]：全文検索サービス等のため、サーバーにある情報を再帰的に取得するユーザーエージェント。

自動巡回ツール：ユーザーが、主に通信料金の節約や更新された情報の検出を行うために使用するユーザーエージェント。オフラインブラウザやプリフェッチャとも呼ばれる。

2.2 製品名の特定

本稿にて提案する手法では、まずログ中のUser-Agent文字列に基づき、ユーザーエージェントを製品別にグループ化する。これは、User-Agent文字列をそのまま製品名として統計等をとっても、種類が多すぎて理解が非常に困難になるためである。

User-Agentヘッダの形式は一応HTTPにて

<製品名>[/*<バージョン番号>*] ... (*<注釈>*)

と定められており、基本的には上記の*<製品名>*を取り出す処理を行えばよい。

しかし、現実に使われているユーザーエージェントはこれに従っていない例も多い。このため、実際にはUser-Agent文字列から製品名への変換規則は、事例に応じて個別に作成する必要がある。また、User-Agent文字列をそもそも生成しない、あるいは生成しているが他の製品と区別できないといったユーザーエージェントも存在する。このような場合には、アクセス元のアドレスやUser-Agent

以外のHTTPヘッダといった情報も利用して製品名を割り出す必要がある。本稿では、この変換規則の作成が最も手間がかかっている部分である。

3.2 特性の識別

特定の製品を用いたアクセスだけを識別したいと言った要求に対しては3.1の段階でも十分であるが、情報の利用のされ方(露出度、意図しない利用)までを知りたい場合には、さらにユーザーエージェントの特性別に分類することが必要である。

製品名からこのような特性別に分類するためには、その製品がユーザーエージェントとしてどのような機能を持っているかを調査し、製品名から特性分類への変換規則を作成しておく必要がある。

この変換規則も、自動生成する手段は現在は存在しないため、様々な製品が新しく発表・使用される度に手動で追加・更新していく必要がある。

4．評価

原データとして、筆者が管理しているHTTPサーバーの、1998年7月～1999年7月の１年間のアクセスログ（合計アクセス数831485,User-Agent文字列は6160種類）を用いた。

第１段階分類としては(1)単にUser-Agent文字列の先頭を切り出す、(2)変換規則を用意し§３で述べた方法を適用する、の２通りについて行い、その上にさらに第２段階の分類を行った。(2)のために用意した変換規則は100種類強、第２段階の製品名→特性分類の対応は370通りほどを用意した。最終的に得られたユーザーエージェントの特性分類結果を図２．に示す。

グラフにてブラウザの次位の大部分を占めている「自動巡回ツール」「検索ロボット」型のユーザーエージェントにより取得された情報は、実際にユーザーが閲覧した可能性は小さいことが予想される。従って、全体の傾向としてはアクセスされた情報の４割は実際には閲覧されていない可能性があると言える。

次に、(1)と(2)の各手法による分類結果の違いに付いて考察する。分類の割合としてはブラウザに関して数パーセントの違いを得た。１日あたりのアクセス数に換算すると、先頭切り出し方式と比べ60アクセス弱/日が「ブラウザに見えるが実際はブラウザではない」と判定されたことになる。

この中には単純なUser-Agent文字列の切り出しでは判定できない自動巡回ツールや自動更新型プロクシサーバーが多く含まれている。特に情報の露出度を予測する場合は、本稿で用いた方法によってより正確なデータが得られることがわかる。

なお、検索ロボットに関しては多くの場合、サーバー側で挙動を制御する設定[5]に対応させるため、(1)と(2)で大きな違いは通常は観測されない。しかし、変換規則を適用することによって、単純な方法では検出の困難であった「メールアドレス採集ロボット」「画像利用監視ロボット」といった、頻度は低いが情報の利用のし方に非常に特徴のあるものを分離することも可能となった。これらの特定は情報の保護と言う観点からも有効であると言える。

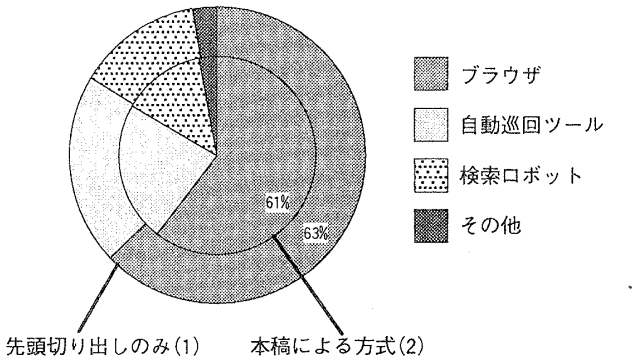

図２．本稿の方式によるユーザーエージェント特性識別の効果

5．まとめ

ユーザーエージェントを特定するための方法として、HTTPサーバーのアクセスログを２段階に分けて解析する手法を示した。この手法により、数値的には数％の違いを得ることが出来た。また、従来は検出の難しかった特徴的なユーザーエージェントを特定することが可能となった。

本稿にて使用した手法はアクセスログの時間軸方向の解析は行っていない。今後はアクセス列を考慮に入れてより精度と手間を軽減する手法を模索してみたい。

A．参考文献

[1] Roy Fielding et al, "*Hypertext Transfer Protocol -- HTTP/1.1*", RFC2616, June 1999

[2] T. Berners-Lee et al., "*Uniform Resource Locators (URL)*" ,RFC1738,December 1994

[3] WWW Consortium, "*Common Log Format*", *http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format*

[4] Martin Koster, "*The Web Robots FAQ*", *http://info.webcrawler.com/mak/projects/robots/faq.html*

[5] Martin Koster, "*A Standard for Robot Exclusion*", *http://info.webcrawler.com/mak/projects/robots/norobots.html*

[6] 河辺岳人,「User-Agentについてのたわごと」,*http://www.dais.is.tohoku.ac.jp/logs/agentgripes.html*