第 3 者認証を施したクローラと Web サーバによるデータの高信頼収集方式の提案

安島　真也†　　星　徹‡　　手塚　悟‡

†東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻
192-0982　東京都八王子市片倉町 1404-1
‡東京工科大学　コンピュータサイエンス学部
192-0982　東京都八王子市片倉町 1404-1
g21120028b@st.teu.ac.jp　hoshi@stf.teu.ac.jp　tezuka@stf.teu.ac.jp

あらまし　クローラは自動で Web サーバにアクセスし，データを収集するプログラムである．しかし，過剰なアクセスを繰り返して Web サーバに負荷をかける，データ収集を制限する構文を無視する等，悪質なクローラが存在する．また，良質なクローラだとしても，アクセス障害を起こす Web サーバが存在する．安定したデータ収集を保証するためにクローラと Web サーバそれぞれに高い信頼性が求められる．そこで本稿では，高信頼のデータ収集を実現するための手法として，クローラおよびサーバを第 3 者機関で認証する方式を提案する．
キーワード：クローラ，第 3 者認証

Proposal of High Authentic Web Data Collecting Method Between Crawler and Web servers adopting Third Party Certification

Shinya Ajima†　　Tohru Hoshi‡　　Satoru Tezuka‡

† Graduate School of Bionics, Computer and Media Science,
Tokyo University of Technology
1401-1 Katakuramachi, Hachiouji, Tokyo 192-0982, Japan

‡ School of Computer Science，Tokyo University of Technology

1401-1 Katakuramachi, Hachiouji, Tokyo 192-0982, Japan
g21120028b@st.teu.ac.jp　hoshi@stf.teu.ac.jp　tezuka@stf.teu.ac.jp

Abstract Crawler is a program that accesses Web servers and collects web data. But, there are malicious crawlers such as putting load on the Web server by repeating excessive access, or ignoring the syntax to protect the data not to be collected. Moreover, even though well-behaved crawlers are used, there are vulnerable Web servers that cause access failures. In order to guarantee the stable data collecting, not only crawlers but also Web servers are required high authenticity and dependability. To answer this issue, we propose a high reliable authentic collecting method between crawler and Web server adopting third party certification.
Keyword：Crawler, Third Party Certification

1　はじめに

1.1　背景

近年，情報技術の発達と情報通信基盤の普及に伴い，インターネット上には Web ページをはじめ，様々なデータが膨大に散在している．その膨大に散在しているデータの中から特定の種類のデータを的確にかつ大量に収集するとなると人手だけではほぼ不可能である．収集を自動化するツールの１つとしてクローラがある．クローラとはインターネット上に存在する Web サーバに自動でアクセスし，定期的にデータを収集，解析するプログラムである．クローラによるデータ収集をクローリングと呼ぶ．Google や yahoo 等の検索エンジンには欠かせない技術であるが，他にもクローラを活用して事業を展開している企業もある．クローラは今のインターネット社会において欠かせないツールとなっている．

1.2　クローラの課題

これまで様々なオリジナルのクローラが研究開発されてきた．しかしながら，オリジナルのクローラが効率良くデータを収集する高性能な仕様であったとしても，Web サーバが安心して受け入れられる保証はない．アクセス先の Web サーバに障害を起こす仕様の可能性がある．

また，クローラの仕様に問題がなくても，アクセスした Web サーバの仕様に不具合あった場合，Web サーバにアクセス障害が起こる可能性もある．

クローラもしくは Web サーバの仕様の不具合によって，データ収集が目的であるクローラ使用者に攻撃の意図があると見なされる恐れがある．安定したデータ収集を実現するためには，クローラは Web サーバにアクセス障害を起こさない仕様であること，Web サーバはクローラによってアクセス障害が起きない仕様であること，お互いの仕様が共に安心，安全なデータ収集を実現するという良質であることを証明する高い信頼性が求められる．

そこで本稿では高信頼のデータ収集を実現するための手法としてクローラおよび Web サーバを第 3 者機関で認証する方式を提案する．

2　関連研究

マルチエージェントクローラ[1]は，インターネット上から非行逸脱傾向が高い有害ユーを発見するために研究開発された．各ユーザが管理する個人領域と個人領域間のリンク関係を収集し，有害ユーザを発見するものである．評価した結果，従来のクローラより効率良く有害ユーザを収集できることが立証されたと述べられている．

パイプライン型クローラ[2]は，任意のモジュールの変更，追加，削除が可能であり，かつ，モジュール間データをリアルタイムストリームとして受信可能であることの条件を兼ね備え，リアルタイム性が高い非構造データを収集するために研究開発された．一般的なクローラはモジュールの変更，追加，削除をするといった作業を行った際，他のモジュールに影響が及ぶ可能性がある．しかし，開発されたクローラは全てのモジュールを完全独立にしたため，ある 1 つのモジュールに対して，変更，追加，削除といった作業を行っても他のモジュールに影響を与えないことを保証したと述べられている．

2 つの関連研究を上記で述べたが，両方のクローラの仕様は高性能な可能性がある．しかしながら，Web サーバが安心して受け入れられるクローラであることの保証はない．

3　クローラが関連した事例

3.1　不規則な動作をするクローラ

通常のクローラは Web サーバ管理者が設定したデータ収集を制限する構文に従う．平成 21 年度著作権法改正ポイント[4]にもインターネット情報検索サービス事業者として満たすべき基準の 1 つとして，「情報検索サービス事業者がクローリングすることについてサイト管理者による禁止措置が取られた情報を収集しないこと」と挙げている．代表的なものは robotx.txt とメタタグがある．robots.txt はクローラに対する命令を記述したファイルであり，Web サイトのトップの階層に設定する．Google のクローラに対して/cgi-bin の下のファイルを検索させない記述例を図 1 に示す．

```
User-agent:Googlebot
Disallow:/cgi-bin
```

図 1　robots.txt の記述例

メタタグは各 HTML 内の<head>と</head>の間にクローラに対する命令を記述する．検索データベースへの登録禁止およびこのページに含まれるリンクをたどることを禁止にする記述例を図 2 に示す．

```
<metaname="robots"content=
                "noindex,nofollow">
```

図 2　メタタグの記述例

しかし，すべてのクローラが上記の構文に対応するわけではない．構文を無視する仕様等をクローラに施せば，Web サーバは対応できず，受け入れざるを得なくなる．2003 年，我が国の一部の Web サイトに某国のクローラが Dos(Denial of Service)攻撃並の訪問を繰返した事例がある．対応策として Web サーバ管理者らはクローラの IP アドレスを拒否する措置をとった[3].

3.2　アクセス障害を起こす Web サーバ

2010 年，あるユーザが自作クローラを某市立中央図書館の蔵書システムにアクセスさせた際，蔵書システムにアクセス障害が発生した．クローラを作成したユーザは業務妨害で逮捕された[5]．しかし，調査した結果，作成されたクローラは一般的なクローラと同等の性能があることが判明し，図書館の蔵書システムに不具合があることを指摘された．図 3 に当時の図書館の蔵書システムを示す．

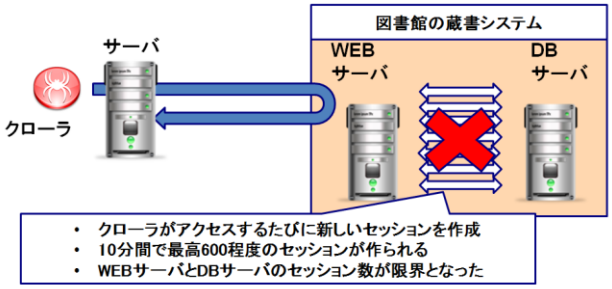

図 3　図書館の蔵書システム

図書館の蔵書システムはアクセスがある度に Web サーバとデータベースサーバの間にセッションが発生し，アクセスが切れても一定時間セッションを保ち続ける仕様であった．そのため自作したクローラがアクセスする度にセッションが作られ，10 分間で最大 600 セッションと限界に達し，アクセス障害が起きた．この事件後，様々な場所で議論され対策案が出たが，Web サーバがアクセス障害を起こした場合の対応，システム管理者の教育等，ほとんどが Web サーバ側の対策案であり，クローラへの対策案は議論されていない．

以上，2 つの事例より，クローラ使用時における課題を下記に示す．

(1) Web サーバ対応できない仕様のクローラが存在する．(図 4)

図 4　Web サーバにアクセス障害を起こすクローラ

(2) 一般的なクローラと同等の性能にも関わらず，アクセスしたことによって障害を起こす Web サーバが存在する．(図 5)

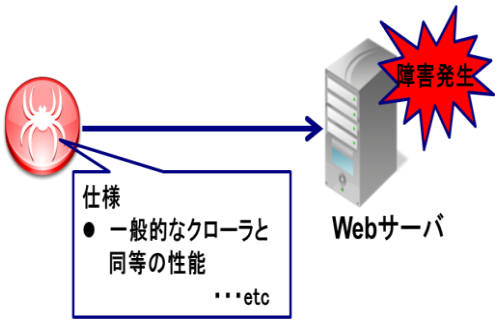

図 5　良質な仕様のクローラによってアクセス障害を起こす Web サーバ

上記の課題を解決するためにはクローラは Web サーバにアクセス障害を起こさない仕様であること，Web サーバはクローラのアクセスによってアクセス障害を起こさない仕様であること，お互いの仕様が安心，安全なデータ収集を実現する良質であることを証明する高い信頼性が求められる．

そこで本稿では，クローラおよび Web サーバの仕様が良質であることを証明する手法を提案する．

4　提案手法

4.1　システムの概要

図 6 に提案手法のシステム概要を示す．

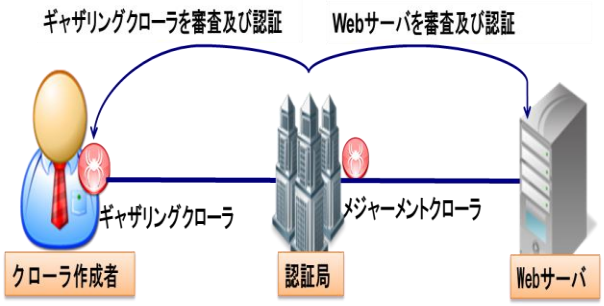

図 6　システム概要

クローラおよび Web サーバそれぞれの仕様を審査および認証する組織として，双方間に第 3 者機関を設置する．第 3 者機関が審査することでクローラ作成者は作成したクローラが Web サーバにアクセス障害起こさず，データ収集ができる仕様であることを確認できる．また，Web サーバ管理者は構築した Web サーバがクローラのアクセス耐えられる仕様であると確認できる．

Web サーバの審査はクローラを使用する．実際にクローラでアクセスすることで，Web サーバの耐久度等を審査することができる．

本稿では第 3 者機関を認証局，認証局で認証するクローラをギャザリングクローラ，Web サーバを審査するクローラをメジャーメントクローラと呼ぶ．

4.2　実現すべき項目

ギャザリングクローラおよびWebサーバの間に安心，安全な環境を実現するために，表 1 に提案手法で満たすべき項目を示す．

表 1　提案手法が満たすべき項目

ギャザリングクローラ	発信元の証明
	仕様が良質であることの証明
	認証局で認証後，仕様が変更されていないことを証明
メジャーメントクローラ	発信元の証明
	仕様が良質であることの証明
Web サーバ	ギャザリングクローラのアクセスに耐えられる仕様であることを証明
	認証されていないギャザリングクローラの拒否

(1) ギャザリングクローラ

・発信元の証明

　ギャザリングクローラが適切な場所から発信していることを Web サーバが確認できるようにする.

・ギャザリングクローラの仕様が良質であることを証明

　アクセスの回数は適度である，データ収集を制限する構文に従う等，使用されるギャザリングクローラの仕様が良質であることを証明できる.

・認証局で認証後，ギャザリングクローラの仕様が変更されていないことを証明

　データ収集で使用されるギャザリングクローラの仕様が変更されていないことを証明できる.

(2) メジャーメントクローラ

・発信元の証明

　メジャーメントクローラが適切な場所から発信していることを Web サーバが確認できるようにする.

・メジャーメントクローラの仕様が良質であることを証明

　Web サーバを審査するためにアクセスしたメジャーメントクローラの仕様が良質であることを証明できる.

(3)Web サーバ

・ギャザリングクローラのアクセスに耐えられる仕様であることを証明

　ギャザリングクローラのアクセスによって障害が起きない Web サーバであることを証明できる.

・認証されていないギャザリングクローラのアクセスを拒否

　Web サーバは認証局で認証されていないギャザリングクローラのアクセスを拒否することができる.

4.3　ギャザリングクローラと Web サーバの認証

図 7 にギャザリングクローラおよび Web サーバの認証完了までのフローを示す. Web サーバがギャザリングクローラおよびメジャーメントクローラの発信元を確認できるようにするために，電子署名を用いる.

1.クローラ作成者は使用するギャザリングクローラを認証局に申請する.

2.認証局は申請されたギャザリングクローラに対し，アクセスの頻度やデータ収集を制限する構文に従うか等,様々な審査を行う. 審査後，認証局は申請されたギャザリングクローラの仕様は良質であることを認証し, 保持する.

3.作成者にギャザリングクローラの使用許可の通知をする.

4.Web サーバは認証局に審査を依頼する.

5.依頼を受け取った認証局は，事前に公開鍵を Web サーバに送信する.

6.認証局の秘密鍵でメジャーメントクローラに電子署名を付与する．そして Web サーバにアクセスし，審査する.

7.Web サーバはメジャーメントクローラの電子署名を公開鍵で復号する.

8.審査後, 認証局は申請された Web サーバは良質な仕様であることを認証し，審査結果を Web サーバに通知する．また，Web サ

ーバの仕様が良質であることを保証する証明書を発行する．

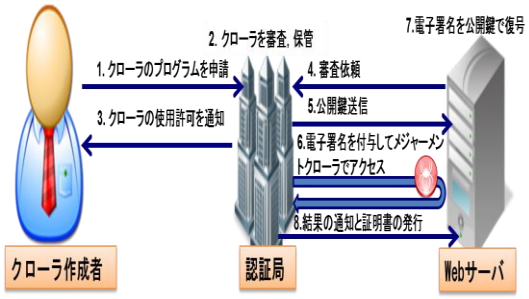

図 7　申請から審査完了までのフロー

4.4　データ収集開始

図 8 にデータ収集のフローを示す．認証されたギャザリングクローラの仕様が変更されていないことを証明するために，認証局からギャザリングクローラを発信させる．

1.クローラ作成者は，認証局にデータ収集の申請をする．

2.申請を受け取った認証局は作成者のギャザリングクローラを起動する．そして認証局の秘密鍵で電子証明をギャザリングクローラに付与し，証明書を保持している Web サーバのみアクセスする．

3.Web サーバは審査時に送られた認証局の公開鍵でギャザリングクローラの電子署名を復号する．

4.収集完了後，収集したデータをクローラ作成者に送信する．

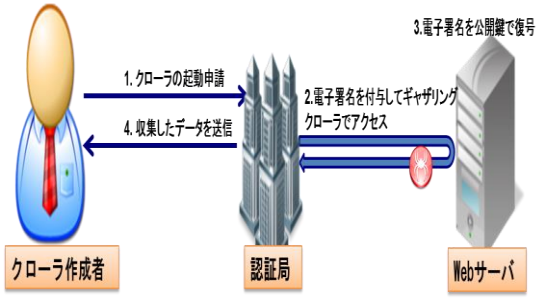

図 8　データ収集フロー

5　まとめ

本稿では高信頼のデータ収集を実現するための手法としてギャザリングクローラおよび Web サーバを第 3 者機関である認証局で認証する方式を提案した．ギャザリングクローラはアクセスした Web サーバに障害を起こさない仕様であること，Web サーバはクローラのアクセスによってアクセス障害を起こさない仕様であることを，それぞれを認証局が証明することで，安心で安全なデータ収集が可能になると考えた．今後の予定として，提案システムが正常に作動するかを確認し，有用性を評価する．また，今後の課題として，メジャーメントクローラの審査方法や Web サイト閲覧者になりすましたクローラへの対策等を検討していく．

参考文献

[1]中村健二，田中成典，北野光一，寺口敏生，大谷和史，”マルチエージェントクローラを用いた有害ユーザの効率的発見手法”，情報処理学会論文誌，Vol53,No.1(2012)

[2] 打田研二，上田高徳，山名早人，“カスタマイズ性とリアルタイムなデータ提供を考慮したクローラの設計と実装”，データ工学と情報マネジメントに関するフォーラム 2012

[3] ジューベー株式会社，”クローラが招く問題 ”http://jubei.co.jp/crawling3.html ，2013/06 参照

[4] 平成 21 年度著作権法改正ポイント，http://www.meti.go.jp/policy/it_policy/daikoukai/igvp/index/h22_report/sub/06.pdf ，2013/06 参照

[5] 日本図書館協会， http://www.jla.or.jp/portals/0/html/jiyu/okazaki201103.html，2013/06 参照