安全で信頼性のあるクローリング環境の提案に関する研究

谷澤 卓† 髙島 麻衣‡ 陳 崇仁‡
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
g211403364@edu.teu.ac.jp g211402374@edu.teu.ac.jp g211301528@edu.teu.ac.jp
hoshi@stf.teu.ac.jp tezuka@stf.teu.ac.jp

あらまし クローラとは，検索エンジン等で用いられるWeb上の情報を自動収集するプログラムである．しかし，クローラは使い方を1つ誤ると，DoS（Denial of Service）攻撃にみなされてしまう可能性がある．またクローラがルールを守っていても，Webサーバの構築環境の問題で，クローラがWebサーバに攻撃したと疑われる場合がある．このような事態を起こさせないために，クローラとWebサーバの両方で信頼性を構築する必要がある．この信頼性を確保するために，第三者機関による安全性が証明されたクローラ及びWebサーバを活用する．本稿では，第三者機関から発行された電子証明書の利用による安全で信頼性のあるクローリング環境を提案する．

Research on the proposal of a safe and reliable crawling environment

Taku Yazawa† Mai Takashima‡ Takahito Chen‡
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

Abstract The crawler is a program used by search engines for automatically collecting on Web. However, the crawler may be regarded as DoS(Denial of Service) attack once its usage is mistaken. In addition, even if it follows the rule, it may be suspected that the crawler attacked on the problem of the construction environment of a Web server. In order not to cause such a situation, it is necessary to build reliability in both the crawler and the Web server. To ensure reliability and safety, we take advantage of crawler and Web server whose safety is certified by a third party. In this paper, I suggest a safe and reliable crawling environment, with a digital certification, which is published by a third party.

1　はじめに

1.1　背景

現在日本国内では企業情報を活用した基盤の整備が進んでいる. 企業情報とは, 企業の日本語や英字の正式名称, 所在地, URL 等の情報である. 現在, JIPDEC がサービスを行っている ROBINS(Reference Of Business Identity for Networked Society)はその基盤の1つである[1]. ROBINS は, 企業が情報を提供し, 第三者が内容を確認することが可能なサービスであるが, このサービスに登録されている企業数が少ないという問題がある. 手動で情報を収集するとなると手間やコストがかかるため, 収集方法の 1 つとして挙げられるのがクローラである. しかし, クローラは使い方を 1 つ間違えるとサイバー攻撃にみなされてしまう可能性がある.

例えば 2010 年に愛知県岡崎市立中央図書館の利用者が図書館のデータベースにクローリングした際に DoS（Denial of Service）攻撃と間違われて逮捕されるという事件が発生した[3][4]. 利用者には攻撃の意志はなく, クローラも特に悪質なものではなく, 原因は図書館側のシステムにあった.

1.2　課題

現在クローラを運用するにあたっていくつか問題が挙げられる. 1つ目の問題は悪質なクローラの存在である. 本稿で定義する悪質なクローラとは, ①必要以上に Web サーバへアクセスをして Web サーバへ必要以上の負荷をかけるクローラ, ②クローラ自身（以後, User Agent と呼ぶ）の成り済ましを行うクローラ, ③Robots.txt を無視したクローラを指す. Robots.txt とは, Web をクローリングするクローラによるサイトへのアクセスを制限するファイルである. クローラはサイトにアクセスする際にトップページに Robots.txt が存在するか確認し, クローラは Robots.txt がある場合記述されている命令に従いクローリングを行う. しかし, Robots.txt には強制力がないため, Robots.txt を無視する悪質なクローラが存在する.

もう 1 つの問題はクローラのアクセスを攻撃とみなされてしまう可能性である. クローラが Web サーバへ必要以上のアクセスをしなく, Robots.txt のルールを尊守し, User Agent も自分のものを使用しているといった3つの条件を満たしていても, クローリングすることでクローラが悪質と誤認されてしまうケースも存在する. これは Web サーバ側の不具合の可能性があるため, Web サーバ側のスペックやシステムの問題で, クローラが DoS 攻撃をしているのではないかと疑われるケースがある.

1.3　目的

クローラを安全に利用するため, クローラ利用者とクローリングされる Web サーバが互いに信頼可能なクローリング方式が必要である. 本稿ではクローラにはデータの収集を行うギャザリングクローラ, 標準的な頻度のクローリングを対象サーバに行い, サーバに異常が発生しないかの評価を行うメジャーメントクローラを定義する. そして, 本稿では, ギャザリングクローラに関して, 第三者機関を使ってギャザリングクローラ利用者とクローリングされる Web サーバが互いに安全で信頼性のあるクローリング環境を提案する.

2　関連研究

第三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[2]はギャザリングクローラ及び Web サーバを第三者機関である認証局で認証することで信頼性を確保した. 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対してアクセス障害を起こさないことを認証局が証明する.

上記の研究では認証局がユーザの代わりに

Web サーバへクローリングすることで安全性を確保している. しかし, 認証局は電子的な証明書を発行・管理する機関であるため, クローリングの役割を付随させるのは実用的ではない.

この問題点に関して, 本稿ではさらに実用的な方式を提案する.

3 技術要素

3.1 クローラ[5]

クローラとは自動的かつ周期的に Web ページから情報を収集するプログラムである. ウェブスパイダー, 検索ボットとも呼ばれ, 主に全文検索型サーチエンジンの検索データベースを作成するために Web を周回している. クローラはデータを Web から収集する際, HTTP(Hyper Text Transfer Protocol), または HTPS(Hyper Text Transfer Protocol Secure)プロトコルを利用する. 収集データはクローラ毎に様々だが, テキストファイル, PDF ファイル, 動画ファイル等がある. HTTP プロトコルでは, リクエストに対してレスポンスを返す. レスポンスにはステータスコードというものが含まれており, そのステータスコードによりクローリングの許可, 拒否を振り分ける. 基本的には 200 を返し, URL が変更されていたら 301, クライアントエラーであれば 4XX, サーバエラーであれば 5XX が返される.

3.2 認証局[5]

認証局(Certificate Authority)とは電子的な証明書(公開鍵証明書)を発行し管理する機関である. 認証局の役割は大きく2つ存在する. 1つ目は電子証明書の発行をすることである. 2つ目は電子証明書や秘密鍵の失効をすることである. 失効を行う理由として, 証明書の有効期限切れと, 有効期限内であっても電子証明書の所有者が自分の秘密鍵を紛失や盗難にあってしまい, 成り済まされてしまう可能性があるからである.

3.3 電子署名[5]

電子署名とは, 電子文書に対して行われる電磁的な署名である. 紙媒体の署名と同様, 電子署名の役割を果たすために完全性の保証, 責任の明示, 以上2つの要件を満たしている必要がある. 署名により改竄検知が可能になるため改竄防止, 改竄することが不可能となるため完全性が保障される. 責任の明示は誰が署名したか証明することができるため, 署名者の責任を明示し, 否認を防止することが可能である. 電子署名の役割・効果を表1に示す.

表1 電子署名の役割

電子署名の役割	効果
完全性の保証	改竄防止
	非改竄証明
責任の明示	否認防止
	文責表明

3.4 電子署名の役割[5]

署名の責任を明示するために, 公開鍵に対する秘密鍵が誰のものか証明する公開鍵証明書が発行される. これにより署名者の特定が可能になる.

電子署名は認証証明の役割を果たす. 認証は, 本人性を確認すること. 証明は, 署名と連携し, 誰がどのような署名をしたかを署名することで可能になる. 電子署名の使い方を表2に示す.

表2 電子署名の使い方

	認証	証明
意味	本人性の確認	資格・属性の証明
用途	アクセス認証	署名・否認防止
対象者	一般人	権限者

3.5 ハッシュ関数[5]

ハッシュ関数とは, 一方向性のためハッシュ値から元ビット列に戻すことは不可能であり, 異なるビット列から同じハッシュ値を出すことは不可能なため改竄防止等の理由で利用される.

3.6　Nutch[6]

NutchとはLucene（全文検索エンジン）のサブプロジェクトとして開発された．JakartaLuceneをベースにクローラや検索エンジンとしてインターフェースも含めたパッケージとしたオープンソースである．Webページのリンクをたどりながら情報を収集し，スコアをつけ，全文検索用のインデックスを作成する．NutchはJavaで作られており，本稿ではNutch2.2.1を使用している．

4　研究目的

安全で信頼性のあるクローリング環境を実現するにあたり以下の要件が必要となる．

・クローラ利用者が安心してクローリングをするためには自分自身のクローラが安全でルールに則っているか
・クローリングをするWebサーバは適切な負荷をかけても問題がない環境になっているか確認し，クローリングされるWebサーバ側もクローリングを実行してくるクローラは安全なものなのか
・Webサーバは一般的なクローラのアクセスに耐えられる耐久力があるのか確認し，安全なクローラのみにクローリングされること

本稿では，クローラ利用者とクローリングされるWebサーバが安全で信頼性のあるクローリング環境構築を提案する．

5　提案手法

5.1　提案手法の概要

クローリングする際にサイバー攻撃と誤認されるのを防止のするため，第三者機関（認証局）を利用してクローラの信頼性のある提案手法を下記に示す．概要を図 1 に示す．また，前提条件を以下に記述する．

・各通信はSSL（Secure Sockets Layer）通信
・ギャザリングクローラおよびWebサーバroot証明書をインストール済み
・ギャザリングクローラおよびWebサーバは秘密鍵・公開鍵を作成済み
・認証局がWebサーバのメジャーメントを行う
・ハッシュ関数はSHA-256を使用
・クローラはNutch2.2.1を使用

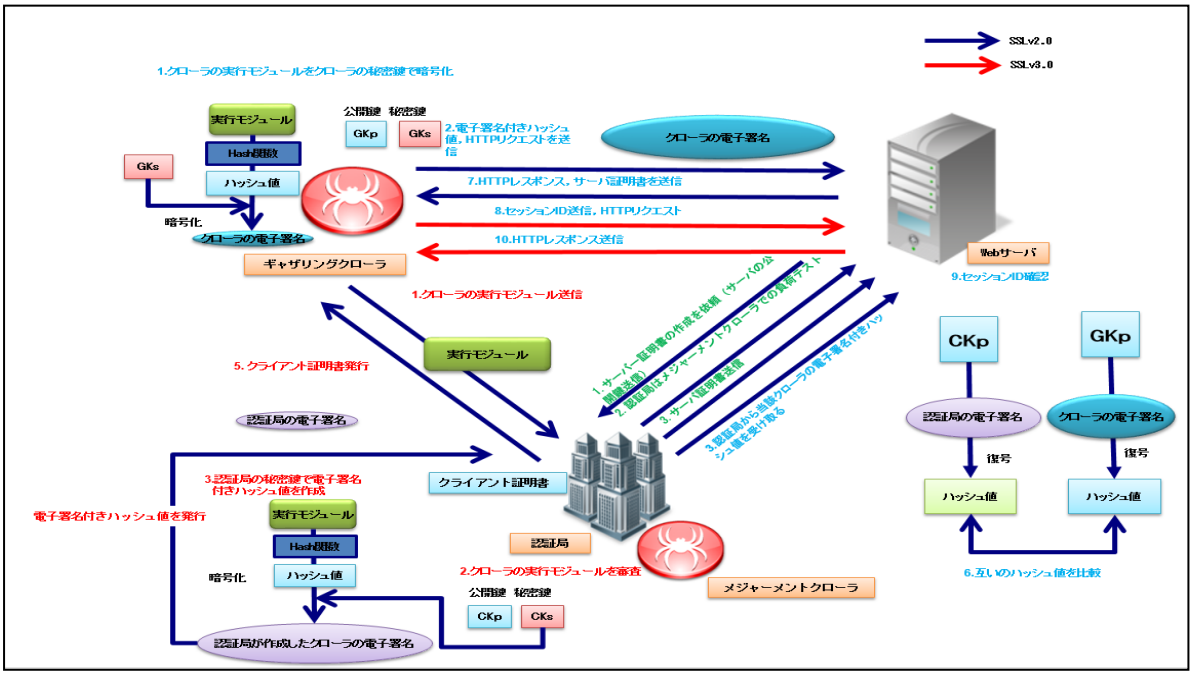

図 1　提案手法の概要図

提案手法では通信を全て SSL 通信で行う. 通信を暗号化することで, 第三者からの盗聴, 改ざんを防ぎ, 通信の信頼性を保持する必要があるためである. 図1では SSLv2.0, SSLv3.0 という 2 種類の SSL 通信を利用している. SSLv2.0 はサーバ認証, セッションの暗号化が必要であり, SSLv3.0 はサーバ認証, ユーザ認証, セッションの暗号化が必要. SSLv3.0 の方が脆弱性は少ない. 図1のフェーズ毎の説明を以下に記述する.

5.2 ギャザリングクローラの審査および認証フェーズ概要

ギャザリングクローラの審査および認証フェーズについて説明を行う. ここで定義する GKp,GKs はギャザリングクローラがインストールされているクライアントの公開鍵, 秘密鍵のことである. ギャザリングクローラの審査および認証フェーズとはギャザリングクローラを認証局に登録するフェーズである. 一連の流れを図 2 に記述する.

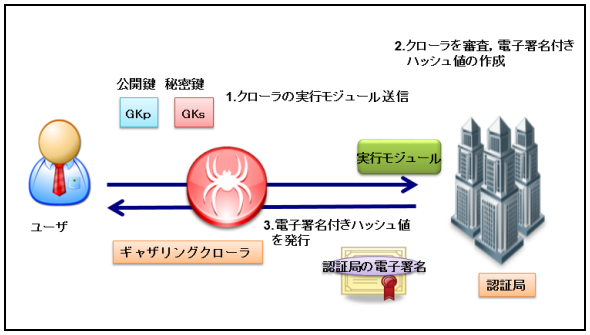

図 2 ギャザリングクローラの審査および認証フェーズの概要図

ギャザリングクローラを作成したユーザは初めに認証局へクライアント証明書を発行してもらわなくてはならない. 手順を以下に示す.

・1. 初めに認証局へクローラの実行モジュール送信を行う.
・2. それを元に認証局はクローラのプログラムを審査, 電子証明書の作成を行う.
・3. 1,2 の手順を経て初めてギャザリングクローラにクライアント証明書発行が行われる.

5.3 ギャザリングクローラの審査および認証フェーズ概要

図 3 はギャザリングクローラの審査及び認証フェーズのデータフローである. ここで定義する CKp, CKs は認証局のサーバが持っている公開鍵, 秘密鍵のことである. 一連の流れを図 3 に示す.

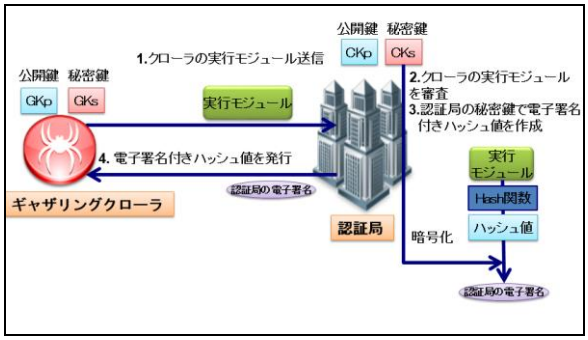

図 3 ギャザリングクローラの審査および認証フェーズのデータフロー

図 3 の手順を以下に示す.

1. ギャザリングクローラが認証局に登録を行う際に, クローラ自身の実行モジュールを認証局に送信する.
2. 認証局はギャザリングクローラから送られてきたデータを参考に信頼できるギャザリングクローラであるか審査を行う.
3. 審査が完了したら実行モジュールでハッシュ値を作成し認証局の秘密鍵で電子署名付きハッシュ値を作成し, 保存する.
4. 全ての工程が終了後, 認証局はギャザリングクローラへ署名付きハッシュ値を発行する.

5.4 Web サーバの審査及び認証フェーズ概要

Web サーバの審査および認証フェーズの説明を行う. Web サーバの審査および認証フェーズとは Web サーバ側の信頼性, システムがある一定水準のクローラからアクセスされても大丈夫なのか確認しサーバ証明書を発行するフェーズである. 概要を図 4 に記述する.

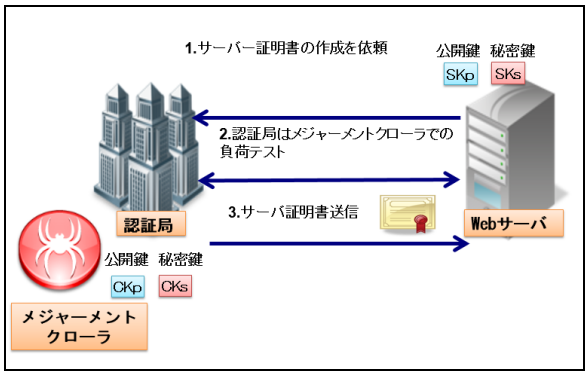

図 4 Web サーバの審査および認証フェーズ概要

図 4 の手順を以下に示す.

1. Web サーバは一般的なクローラのアクセスに耐えられる環境に存在しているのか認証局に証明してもらうためにサーバ証明書の作成依頼を行う.
2. Web サーバの環境を審査するため HTTPS 通信を行い Web サーバの審査を行う.
3. 審査が完了し, 基準に耐えられる環境であれば, サーバ証明書送信をする.

5.5 データ収集フェーズ概要

図 5 はギャザリングクローラがデータ収集を行う際の動きを示したものである.

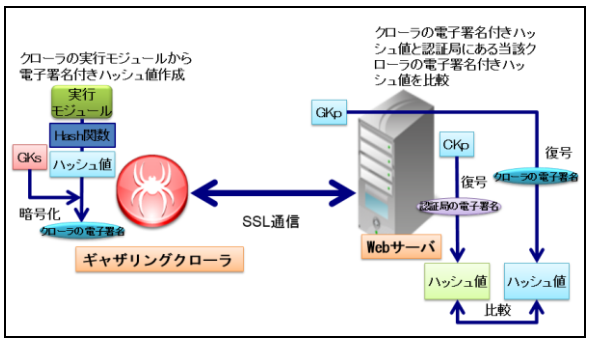

図 5 データ収集フェーズ概要

ギャザリングクローラの審査および認証フェーズ, Web サーバの審査および認証フェーズを行った後, データ収集フェーズを行うことができる. データ収集フェーズではギャザリングクローラと Web サーバ間を SSL 通信でつなぎクローリングを行う.

5.6 データ収集フェーズデータフロー

ギャザリングクローラの審査及び認証フェーズ時から認証局が保持しているギャザリングクローラの電子署名付きハッシュ値と, セッションを張る直前に作成したクローラの電子署名付きハッシュ値を認証局に送信し, そのハッシュ値を比較する. 一致したらクローリングの許可をする. これによりギャザリングクローラの信頼性を維持する. 図 6 はデータ収集フェーズのデータフローである.

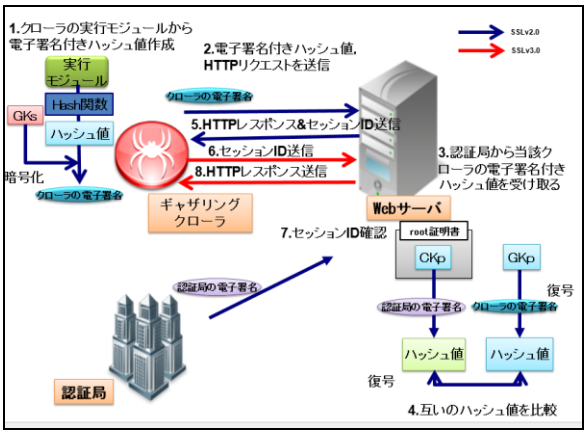

図 6 データ収集フェーズデータフロー

図 6 の手順を以下に示す.

1. 初めて Web サーバへクローリングを行う場合, クローラの実行モジュールから電子署名付きハッシュ値の作成を行う.
2. ギャザリングクローラは電子署名付きハッシュ値を Web サーバに送信する.
3. Web サーバは認証局から当該クローラの電子署名付きハッシュ値を得る.
4. Web サーバは互いのハッシュ値を比較し, 同じであるか確認を行う.

5 ハッシュ値が一致した場合, Web サーバは HTTP レスポンス＆セッション ID を送信.

6 ギャザリングクローラは Web サーバへセッション ID 送信を行う.

7 Web サーバはセッション ID の確認を行う

8 HTTP レスポンス送信をギャザリングクローラへ行う.

6　まとめ

企業情報を手動でデータ収集するには手間やコストがかかる. そこで収集方法の1つとしてクローラが利用される. しかし, クローラの運用にあたっていくつかの問題が挙げられる. 1つが悪質なクローラの存在, もう1つがクローラのアクセスが攻撃とみなされてしまう可能性である. クローラ利用者とクローリングされる Web サーバがお互い信頼できるようにするため, 本稿では新しいクローリング方式を提案した. 今後はシステムを実装し, ROBINS で構築されたデータベースへ企業情報を集積するためのクローリングを行い, 実際に取得した企業情報のデータ数, 認証局への負荷, サーバへの負荷等から本提案方式の有用性について評価を行う. また, 課題として, サーバ・クローラへの電子署名付きハッシュ値等の発行する基準をどのように設けるかの検討も行う.

参考文献

[1] ROBINS
https://robins.jipdec.or.jp/robins/ 2014/7

[2] 安島真也, 星徹, 手塚悟, “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”

[3] Librahack
http://librahack.jp/ 2014/7

[4] 突然、ある技術者が逮捕された：記者が掘り下げた岡崎市立図書館事件
http://astand.asahi.com/magazine/wrnational/special/2011012800004.html 2014/6

[5] 手塚悟,「情報セキュリティの基礎」共立出版

[6] Apach Nutch
http://nutch.apache.org/ 2014/6

[7] 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) 2014/7

[8] 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 2012/6

[9] 著者:竹本 秀樹, 双紙 正和, 宮地 充子 “DoS 攻撃に対する偽造体制をもつ改良パケットマーキング法” 社会団法人情報処理学会研究報告書 2012/12

[10] 著者:長尾 宗胤, 遠山 宏明, 冨澤 眞樹 “パケットフィルタリング機能を搭載した NIC による DoS 攻撃対策” 2013/5

[11] GlobalSign “認証局とは？“
https://jp.globalsign.com/service/knowledge/ca/ 2013/5

[12] JIPDEC “電子署名・認証センター”
http://www.jipdec.or.jp/esac/intro/shikumi.html 2013/5

[13] 著者:次世代電子商取引推進協議会 “電子署名普及に関する活動報告 2008”
http://www.jipdec.or.jp/archives/ecom/results/h20seika/H20results-14.pdf 2013/5

[14] 角田裕, 荒井健二郎, 和泉勇治, 根本義章 “パルス型 DoS 攻撃の被害軽減のためのトランスポート層プロトコルの通信制御に関する検討” 2013/5

[15] 西竜三, 堀良彰, 桜井幸一 “無線通信における物理レイヤ/MAC レイヤへの DoS 攻撃に耐性を有する整合フィルタを用いた符号化方式” 2013/6

[16] Udi Ben-Porat, Anat Bremler-Barr, Hanoch Levy“Vulnerability of Network Mechanisms to Sophisticated DDoS Attacks” 2013/6

[17] 鬼頭利之, 斎藤孝道 “SSL（Secure Socket Layer）のシステムとしての安全性の考察”2014/1

[18] 山田明, 三宅優, 寺邊正大, 橋本和夫 “SSL/TLS で暗号化された Web 通信に対する侵入検知システム” 2013/6