LD-001

サーバ接続環境調査のための低負荷クローリング手法の検証

Evaluation of Low-load Crawler Designed for Server Connection Environment Survey

星野 哲哉† 中平 勝子† 三上 喜貴†
Tetsuya Hoshino Katsuko T. Nakahira Yoshiki Mikami

1. はじめに

OECD の統計によれば，近年，先進国でのブロードバンド普及率は年々向上している [1]．しかし途上国でのブロードバンド普及率は未だに低い．表 1 はアフリカの通信回線帯域を調査したものである [2]．これによると，南アフリカは 45Mbps の帯域を 15 の ISP で分け合っており，スワジランドでは 1 つの ISP 当たり 42Kbps 程度しか使用できない．

表 1　アフリカの通信回線帯域

国名	IXP 名	ISP 数	帯域幅
South Africa	Johannesburg JINX	15	45Mbps
Kenya	Nairobi KIXP	13	6Mbps
Mozambique	Maputo MozIX	7	4Mbps
DRC	Kinshasa IX(KINIX)	4	1Mbps
Nigeria	Ibadan IBIX	2	200Kbps
Tanzania	Dar es Salaam TIX	10	1Mbps
Swaziland	Mbabane SZIX	3	128Kbps
Rwanda	Kigali RINEX	6	400Kbps

一般的に，サーチエンジンのための Web データ収集や，サーバ利用実態の調査，ネットワーク実態調査などを行う際にはインターネットに接続された Web サーバに対する悉皆的なクローリングが必要である．しかし，悉皆的なクローリングは通信回線を著しく圧迫するため，上記のようなブロードバンドが発達していない途上国地域のサーバを調査する際には，相手サーバに大きな負荷がかかることが問題となっている．実際，筆者らがクローリングのためにアクセスした，アフリカの或る Web サイトの管理者から，”I notice that you consider one HTTP request every 5 seconds ’polite’ and ’modest’. This may be true in Japan, but not in Africa .”といった苦情を受けた．この苦情から，途上国地域の通信事情は，我々が想像する以上に厳しいものであることが伺える．そのため，実態調査を行う際にも，対象国の通信事情を考慮する必要があり，より負荷の少ないクローリング方法が必要となっている．

クローリングの際の負荷低減を目指す試みは，フォーカストクローリングの研究に多く，[4] や [5] などで論じられている．[4] では，特定言語で記述された Web ページを選択的に収集するための手法を開発している．また，[5] では，Web 空間のリンク構造の特性を元に，移動した Web ページの移動先を効率よく探索する方法を提案している．

フォーカストクローリングではページ単位のデータを取得することを目的としているのに対して，筆者らの目的とするサーバ単位のデータ取得においては，負荷低減のための戦略も自ずと異なってくる．

筆者らは，これまでの研究において，Web の特性を利用し，サイト外へのリンク情報の収集効率を高く保ちながら，クローリング時の負荷を低減させる手法を開発した [6]．これは，Web サイトを一定の深さに限定して，コンテンツに含まれるリンク情報の抽出を行うクローリング手法である．その際，クロール量を抑えるための手法として，クロールを行った URL の，クエリを削除したものを既読として登録した．これにより，クエリが違うだけの URL は再びクロールされることはない．

しかし，[6] では，URL のクエリを削除したことにより，結果にどのような影響があるかを論じていない．本稿ではその手法の妥当性について検証を行う．

また，上記とは異なる手法によって，サイト外へのリンク情報を効率よく収集し，負荷を低減させる新たな手法を提案する．

2. 研究方法

2.1 検証に使用したデータ

検証に使用したデータは，筆者らが言語間デジタルデバイドを調査する際に，Web 空間から収集を行ったものである．

クローラーには UbiCrawler[7] を用いた．クロールは，表 2 に示す条件で行った．UbiCrawler は，各サーバへの処理対象 URL の割り当てをハッシュを用いて行うことから，どのサーバでも均等にクローリングがなされているとみなし，1 台のみを対象として検証用のデータを抽出した．検証に使用したデータの数を ccTLD(country code Top Level Domain) 別に分類したものが，表 3 である．

表 2　クロールの条件

サーバ数	20 台
スレッド数	各 150
最大深さ	16
最大ページ取得数/ホスト	1,000,000
取得間隔	10 秒

データ収集時期は 2005 年 8 月，対象とした地域はアジア各国 43 カ国の ccTLD を範囲としてクロールしたものである．ページ数は 2,736,309 ページであり，含まれるユニーク URL 数は

† 長岡技術科学大学

表 3　検証に使用したデータ

国名	ccTLD	ページ数	ドメイン数
Israel	il	1,033,067	1,327
Singapore	sg	267,815	218
Thailand	th	229,960	253
Turkey	tr	196,855	505
Indonesia	id	171,678	137
Malaysia	my	121,920	287
Viet Nam	vn	119,865	73
India	in	98,724	205
Iran	ir	89,680	259
Kazakhstan	kz	81,098	117
Oman	om	50,100	7
Uzbekistan	uz	49,179	69
Azerbaijan	az	45,575	23
Jordan	jo	42,631	10
Saudi Arabia	sa	38,883	28
Philippines	ph	32,365	79
UAE	ae	13,262	46
Pakistan	pk	9,986	21
Myanmar	mm	8,425	1
Kyrgyzstan	kg	6,335	72
Kuwait	kw	5,928	6
Bangladesh	bd	4,443	5
Sri Lanka	lk	3,835	12
その他		14,700	131
合計		2,736,309	3,891

10,508,904 である．また，ページ間のリンク総数は 65,956,029 となっている．今回の検証では，主にクエリを削除した場合の影響を検証するため，特に Web ページが多く，Web 空間的にも相応に成熟している [8] と判定された国のうち，.sg（シンガポールの ccTLD）を持つ URL の中から 124 個を選択し，クロールの起点 URL とした．

また，UbiCrawler は，URL とその URL に含まれるリンク情報を出力することが可能であるため，この 2 つの情報を元に，検証用データベースを構築した．このデータベースは擬似的な Web 空間を模しており，URL を問い合わせることによって，その URL に記述されていたリンク情報を返す仕様となっている．問い合わせによって得られたリンク情報を元に，再帰的なクローリングを行うことも可能である．本稿では，同じ母集団に対して複数回クローリングを行う必要があるため，上記のようなデータベースを使用した．

2.2　検証方法

検証方法は，図 1 に示す通りである．

図中の内部 URL とは，起点となる URL と同じドメイン，若しくは，そのサブドメインを有するリンク情報である．外部 URL とは，内部 URL に分類されない URL である．外部 URL は，起点となる URL とは別のドメインを持っており，サーバが別であると認識される．この手法では，外部 URL は収集対象であるため，新たな起点 URL として利用する．（表 4 参照）

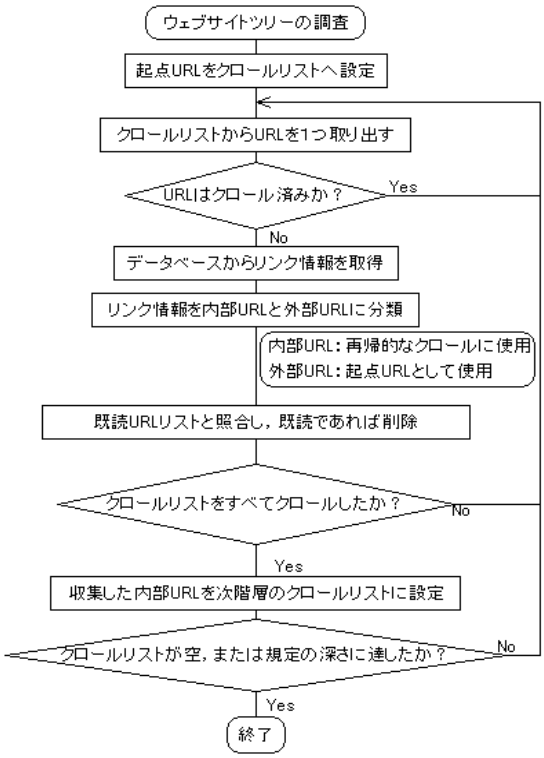

図 1　検証方法

表 4　内部 URL と外部 URL の例示

起点 URL	取得した URL	分類
example.jp	example.jp	内部 URL
example.jp	www.example.jp	内部 URL
www.example.jp	example.jp	外部 URL
www.example.jp	sv1.example.jp	外部 URL

初めに，データベースよりリンク情報を取り出す．

次に，得られたリンク情報を，前述の内部 URL と外部 URL の定義に従って分類する．内部 URL に分類された URL 群は，次階層のクローリングに利用される．また，外部 URL は，この手法における収集対象である．次階層のクローリングリストは，内部 URL から既にクロール済みである URL（以下，既読 URL）を除いた URL 群を利用する．従来方法 [6] では，調査対象サーバへの負荷を軽減するため，既読 URL へ登録する際にクエリを除いた URL を使用していた．しかしこれによって外部 URL 取得数が減少する可能性がある．本稿では，外部 URL 取得数への影響を調査するため，クエリ削除を行わない場合と行う場合の比較を行った．

これ以降，クエリを削除する操作を「クエリ制限」と定義する．「クエリ制限付き」とはクエリを削除する操作を施した結果であり，「クエリ制限無し」とはクエリを削除する操作を行わず，生データをそのまま利用したものである．

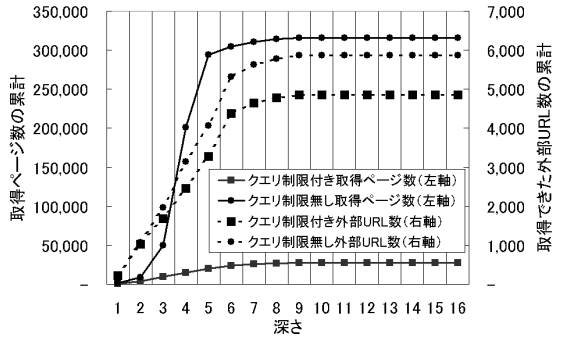

図 2　取得ページ数と外部 URL 数の関係

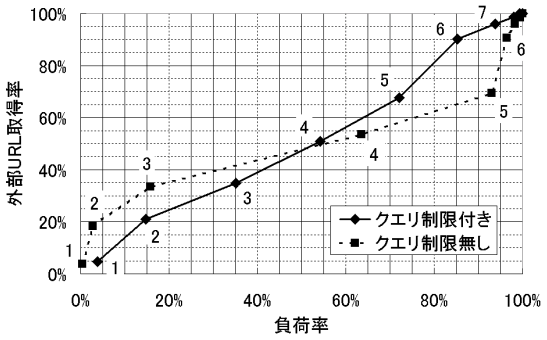

図 3　負荷率と外部 URL 取得率の関係

3.　検証結果・考察

3.1　取得ページ数と外部 URL 数の関係

図 2 は，横軸にサイトツリーの深さ，縦の左軸に取得ページ数の累計，縦の右軸に取得できた外部 URL 数の累計をとって，最大深さを変化させたときに，取得ページ数と，それによって取得できる外部 URL 数がどのように変化するかをグラフ化したものである．図中の実線が取得ページ数を，破線が取得した外部 URL 数を表している．また，クエリ制限付きのグラフは四角のマーカーを，クエリ制限無しのグラフは円形のマーカーを使用している．

この図から，クエリ制限の効果を読み取ることが出来る．例えば，クエリ制限無しの場合，最大深さでの取得ページ数は約 316,000 ページであるのに対し，クエリ制限付きでは，27,500 ページと，クエリ制限無しの場合の取得ページ数に比べて約 9% である．しかし，それによって得られる外部 URL 数は，クエリ制限無しでは約 5,800 件，クエリ制限付きでは約 4,800 件と，約 17% 減少したのみである．即ち，クエリ制限の有無によって外部 URL の取得数に 1,000 件の差が現れたが，この 1,000 件を収集するために，約 280,000 ページのクローリングが新たに必要であると言える．従って，負荷を効率よく抑え，且つ，得られる外部 URL 数への影響を最小限に抑えるためにクエリ制限は有効であると考えられる．

外部 URL の取得数が収束する深さに注目すると，これまでの研究では，深さ 9 において外部 URL の取得数が収束していた．今回の検証でも，取得数の増加は深さ 6 で緩やかになるが，深さ 9 前後でほぼ飽和状態に達した．

3.2　負荷率と外部 URL 取得率の関係

図 3 は，縦軸に外部 URL 取得率を，横軸に負荷率をとって，深さごとにプロットしたものである．実線がクエリ制限付き，破線がクエリ制限無しを示している．ここで外部 URL 取得率と負荷率は，本稿では以下の式で求められる．いずれも分母には深さを無限大としたときの外部 URL 数，ページ数を用いるべきであるが，これは不可能であるため，十分な深さとして D=16 と設定した．

$$\text{外部 } URL \text{ 取得率} = \frac{\text{深さ } n \text{ における取得可能外部 } URL \text{ 数}}{\text{深さ } D \text{ における外部 } URL \text{ 数の総数}}$$

$$\text{負荷率} = \frac{\text{深さ } n \text{ におけるクロール済みページ数}}{\text{深さ } D \text{ におけるページの総数}}$$

例えば深さ 1 で全体の 10% をクロールし，外部 URL 取得数が取得可能総数の 5% 取得できたとすると，最初の点は縦軸が 5%，横軸が 10% の位置にプロットされる．即ち，深さ n におけるクロールの完了度合いと，外部 URL の収集率を示している．また，外部 URL の取得率を下げることによって，どれほどの負荷を低減させることが可能であるかの見当をつけるための指標としても利用することが出来る．

この図で注目すべきは，クエリ制限無しグラフである．このグラフは大きく 3 つの部分に分けることが可能である．まず深さ 1 から 3 までの間に，外部 URL 取得率が大きく上昇している．しかし深さ 3 から 5 までは逆に負荷率の上昇が大きく，深さ 5 以降は再び外部 URL 取得率の大幅な上昇が起きている．

深さ 1 から 3 という比較的浅い階層に外部 URL が集中しているサイトは，個人的に作成されたリンク集やブログサイトであった．対して深さ 5 以降という比較的深い階層に外部 URL が集中しているサイトは，自サイト内のリンクを重視する企業や教育機関のサイトであった．このようなサイト群が混在していたため，図 3 のような変曲点のある屈折したグラフになったと考えられる．

3.3　リンク構造特性に応じた可変サンプリング手法

最大深さ固定クローリングとは異なるアプローチとして，各深さにおけるクロール対象 URL 数そのものを減少させることによって負荷低減を行う手法を提案する．

図 4 は，縦軸に URL の取得率を，横軸にサンプリング率をとって，サンプリング率を変化させた場合における取得 URL 数への影響を示したものである．実線は内部 URL の取得率を，破線は外部 URL の取得率を示している．また，灰色の破線は URL 取得率=サンプリング率を表しており，全てのページが均一にリンク情報を持っていた場合，内部 URL 取得率と外部 URL 取得率はこれに従う．ここでサンプリング率は，深さ $n+1$ でのクロール対象を決定する際に使用される．深さ $n+1$ でのクロール対象は以下の式で決定される．

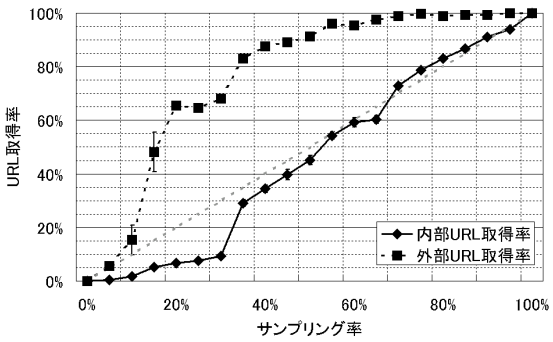

図 4　サンプリング率と URL 取得率の関係

$$\text{深さ } n+1 \text{ のクロール対象} = \text{深さ } n \text{ における総内部 } URL \times \text{サンプリング率}$$

例えば，サンプリング率を 10% とした場合，深さ n にて取得した内部 URL の 10% をランダムに選定し，深さ $n+1$ のクローリングに使用することになる．

また，URL 取得率とは以下の式で表される．

$$URL \text{ 取得率} = \frac{\text{深さ } n \text{ における } URL \text{ 取得数}}{\text{深さ } D \text{ における } URL \text{ の総数}}$$

ここで，サンプリング率を 30% とした場合，クローリングを行った内部 URL は全体の約 10% である．また，その場合に取得出来た外部 URL は全体の約 68% である．即ち，サンプリング率を 30% に設定することにより，深さ D における総外部 URL 数の 68% を取得でき，クローリングにかかる負荷を 10% 程度に抑えることが可能である．さらに，サンプリング率を 55% に設定することで，深さ D における総外部 URL の 95% を取得でき，負荷を約 55% に抑えることが可能となる．

図 4 の場合，内部 URL 取得率が灰色の破線に従い，且つ外部 URL の取得率が十分であるため，サンプリング率は 55% が適当であると判断される．また，URL 取得率の閾値をユーザが決定することにより，適当なサンプリング率を算出することも可能である．

当然ながら，このサンプリング率はクロール対象のリンク構造特性によって変化する．そのため，これを利用するには，クロール対象のリンク構造を事前に分析することが必要である．リンク構造の分析には，クロール対象の全数調査が必要となるため，その時点では，調査対象に最大負荷がかかることになる．しかし，この操作により決定したサンプリング率を利用することにより，その後の調査では負荷を効率よく抑えることが可能となる．さらに，このサンプリング率の作成を定期的に行うことにより，リンク構造特性の変化に柔軟に対応することが可能となる．

4.　おわりに

Web サイトを悉皆的にクローリングする場合，一般的には深さに制限をかけず，全てのリンクを辿らなければならない．しかし本論のように，サーバを補足するためのクローリングでは，深さ 9 までのクローリングにより，十分なサーバ補足率を得ることが可能であることを確認した．また，クエリ制限によって，サーバ補足率の減少を抑えつつ，負荷を低減させることが可能であることを確認した．

しかしながら，Web 利用技術が進歩したことにより，Web 空間におけるコンテンツの形態は刻々と変化している．近年ではブログツールに代表される CMS の発達により，個人が比較的容易に Web サイトを所持・管理することが容易になってきた．そのような CMS 上ではサイドメニューの自動生成によって各ページがサイト内で密接に結びつくことによるサイトツリーの低階層化や，トラックバックによる外部リンクの増大という現象が起こっている．そのため，定期的にリンク構造を調査し，リンク構造の特性に沿ったクローリングを行う手法も提案した．

サーバを単位とするデータ収集を，効率的かつ低負荷で実行することを特徴とする本手法は，デジタルデバイドに関する現状把握の一環として，サーバの所在地などを調査するために開発されたものではあるが，この他にもサーバ技術の動向調査やネットワーク現状調査など，広い応用範囲を持つと考える．

参考文献

[1] OECD Broadband Statistics to June 2006

[2] The Acacia Atlas 2005, The International Development Research Centre

[3] Katsuko T. Nakahira et. al.: ”GeographicLocation of Web Servers under African Domains”, The 15th International World Wide Web Conference, Edinburgh. 2006

[4] 田村孝之 クンワディー・ソンブーンウィワット 喜連川優,”特定言語で記述された Web ページの選択的収集手法とその評価”, 電子情報通信学会論文誌 Vol.89-D No.2 pp.199-209

[5] 澤奈津美 飯田敏成 森崎厚行 ほか,”Web ページ移動先発見のためのクローリング手法の提案”, 情報処理学会 研究報告 2006-DBS-140 (II)

[6] 星野哲哉 中平勝子 三上喜貴,”サーバ接続環境調査のための低負荷クローリング手法の開発”, 情報処理学会全国大会講演論文集，Vol.69 No.1 page.1.495-1.496

[7] http://law.dsi.unimi.it/

[8] 石原直幸 中平勝子 三上喜貴,”Out-degree 分布を用いた Web 利用構造の分析”, 情報処理学会全国大会講演論文集, Vol.69 No.4 page.4.561-4.562