5S-4

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

星野哲哉　　中平勝子　　三上喜貴
長岡技術科学大学

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。

筆者らは、ネットワーク接続を含めたデジタルデバイドに関する実態調査のために、インターネットに接続されたウェブサーバに対する悉皆的なクローリングを行っている[2]。しかし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場合には通信回線に対して多大な負荷がかかることが問題となっている。

本研究では、クローリング時の負荷を低減させるため、ウェブサイトを一定の深さに限定してリンク情報の抽出を行う、準悉皆的な低負荷クローリング手法を開発することを目的とする。

２．調査方法

本研究は、起点となる URL から、同一ドメイン内で辿ることができるページ群を「サイトツリー」と定義する。また、起点となる URL から、リンクを辿る回数を「深さ」と呼ぶ。本手法での処理プロセスは図 1 の通りである。

まず、クロールの起点となる URL から HTML ファイルを取得する。このファイルに含まれるリンク情報を、ドメインを基準として内部 URL と外部 URL（後述）に分類する。以降、内部 URL に分類された URL に対して、HTML ファイルを再帰的に取得する。これをサイトツリーに対し一定の深さだけ繰り返すことにより、リンク情報を準悉皆的に収集する。

２．１．内部 URL と外部 URL の定義

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。これはサイトツリーの定義に外れるため、再帰的なクローリングは行わず、新たな起点 URL として使用する。

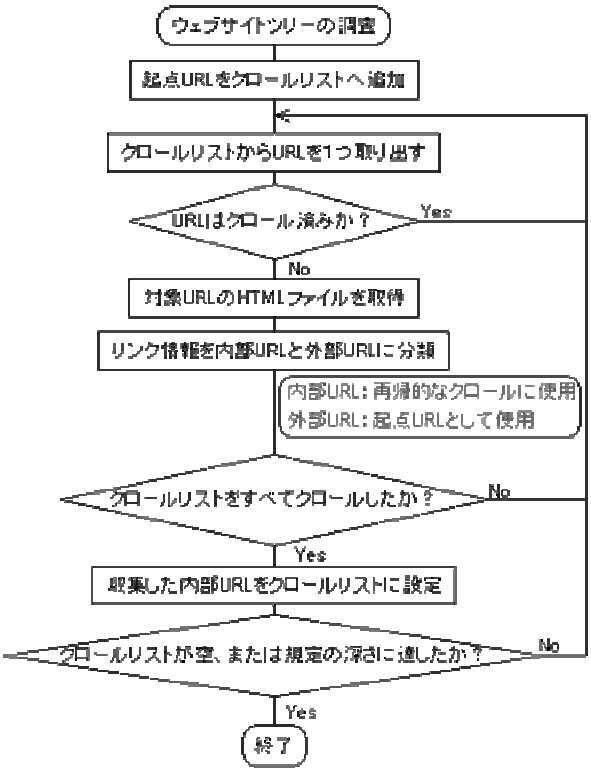

図 1 本手法のフローチャート

２．２．クエリ文字付き URL に対する制限

クエリ文字が付加されている URL では、場合によっては無限に生成され、悉皆的クローリングの際には障害となる可能性がある。そこで本手法では、1 つのスクリプトに対し、1 回のみのクロールに限定することにより、クロール対象の増加を防いだ。例えば、「/index.php?id=2」のクロールを行った場合、これ以降「/index.php」のクロールを一切行わないというものである。

Development of Low-load Crawler to Survey Server Connection Environment
Tetsuya Hoshino, Katsuko T. Nakahira, Yoshiki Mikami
Nagaoka University of Technology

３．調査結果

本研究では、探索する深さを変更することによってクローリングの範囲を限定し、低負荷を実現する。その最適な深さを探索するため、実際にクローリングを行い、深さとそれによって取得できる外部 URL 数の関係を調査した。

調査に使用した、起点となる URL の数は、約 3400 である。これらは、アフリカの ccTLD を持つ URL である。また、サイトの深さを 20 に限定してクローリングを行った。

３．１．取得ページ数と外部 URL 数の関連

図 2 は、縦軸に取得したページ数の累積と、それによって取得できた外部 URL 数の累積を、横軸にサイトの深さをとり、グラフ化したものである。取得したページ数の累計を示すグラフは、深さ 7 を境として傾きが変化することが観察できる。また、深さ 7 以降、深さ 20 になっても収束する様子が見られない。逆に、取得できた外部 URL 数の累計は、深さ 9 でほぼ飽和状態となることが観察できる。

３．２．最大深さ別サイトツリー数の分布

図 3 は、縦軸にサイトツリー数を、横軸にサイトの最大深さをとりグラフ化したものである。深さ 8 以降では、サイトツリーの数は少なく、急速に減少している。

３．３．負荷と取得可能外部 URL 数の関係

図 4 は、縦軸に取得可能 URL 数の割合を、横軸に負荷をとり、グラフ上に表したものである。深さ 20 の時点での負荷の値を 100%とした。この図より、負荷を 20%削減したとしても、取得可能外部 URL 数には全く影響が無いことを示している。

４．結論

言うまでもなく、クロールの負荷と捕捉率とはトレードオフの関係にあり、深さをパラメータとして、ネットワーク接続環境に応じて最適な負荷 - 捕捉率バランスを選択しうる。その目安を与えるものが図 4 である。深さ 8 程度にクローリングを制限することにより、準悉皆的なクロールが、約 80%の負荷で可能となる。また、外部 URL の取得率を 80%まで許容するとしたら、負荷を約 50%低減させることが可能である。図 4 を用いることにより、クローリングの目的と調査対象サーバのネットワーク接続環境に応じたクローリング戦略を選択することができる。また、図 4 は深さを制限したときの捕捉率が与えられているので、全数を捕捉する手がかりも与えている。

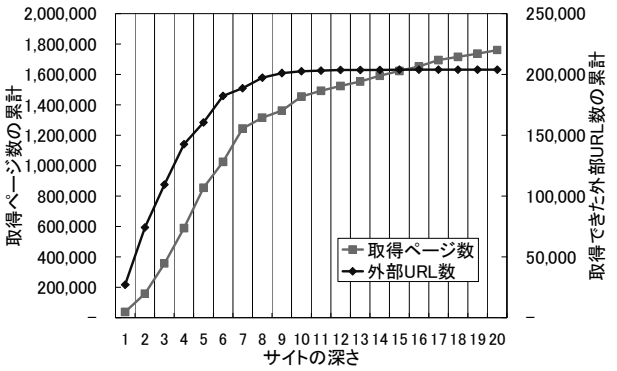

図 2 取得ページ数と外部 URL 数の関係

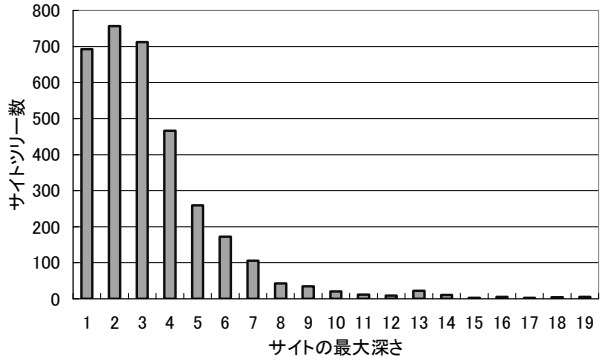

図 3 最大深さ別サイトツリー数の分布

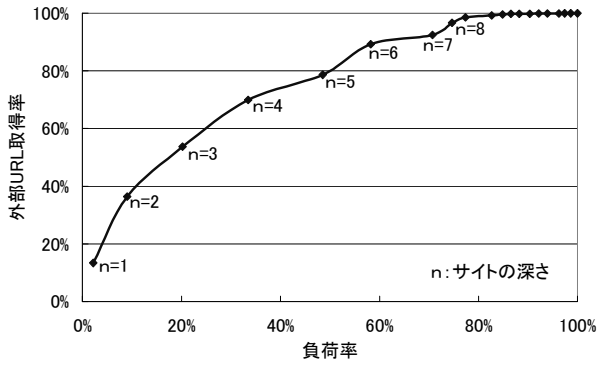

図 4 負荷率と外部 URL 取得率の関係

参考文献

[1]OECD Broadband Statistics to June 2006

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