5G-1

情報収集ロボットの収集特性の解析

竹野 浩 † 能登 信晴 ‡

† NTT サイバーソリューション研究所 ‡ NTT サイバースペース研究所

1 はじめに

インターネット上の Web ページの総数は増え続けている．このため，「情報収集ロボット」と呼ばれるシステムを用いて Web ページを自動的に収集してインデックスを作成し，検索サービスを提供するロボット型検索サービスは，より大量の Web ページを検索対象とすることが求められている．これには検索システムの高速化だけではなく，情報収集ロボットの高速化も不可欠である．Web ページを高速に収集するため，複数の計算機で構成される情報収集ロボットの開発を行い，収集特性の解析を行ったので報告する．

2 情報収集ロボットの性能の解析

2.1 性能指標の設定と制約条件の導出

情報収集ロボットは高速である方が望ましい．しかし情報収集ロボットによる Web ページの収集は Web サーバに大きな負荷を与えるため，1 Web サーバあたりの収集速度は抑制する必要がある．そのため，収集間隔 T_{wait}，最大同時接続数 $N_{connect}$ を設定し，収集終了後 T_{wait} 以内には同じ Web サーバに接続しないものとし，同じサーバに対して $N_{connect}$ を越えた接続を同時に行わないものとする．単位時間に収集する Web ページの数を収集速度 V とすると，性能の良い情報収集ロボットとは，V, T_{wait} の値が大きく，$N_{connect}$ の値が小さいものであると言うことができる．本報告では，$N_{connect}$ は，最小の値である 1 を前提として，情報収集ロボットの性能は (V, T_{wait}) で表現するものとする．

情報収集ロボットの動作は，1 つの Web ページを収集するのに要する時間の平均値を $\bar{t}$ とすると，$V = 1/(\bar{t} + T_{wait})$ となる．情報収集ロボットで制御可能なパラメータは T_{wait} のみであるが，前節で定義した性能指標によりこの値を小さくすることは望ましくない．従って V を大きくするには同時に複数の Web ページを収集する必要がある (図 1)．同時接続数を N とすると，$V = N/(\bar{t} + T)$ となり，N に比例して性能が向上するが，$N_{connect} = 1$ であるため，同時接続数の最大値 N_{max} は，Web サーバの総数を越えることはできない．また，図 1 における接続回線の容量は有限であるため，回線容量を C，Web ページの平均サイズを $\bar{s}$ とすると，$\bar{s}V < C$ である必要がある．

2.2 予備実験

現実の情報収集ロボットの開発において，上記制約のどちらが重要であるかを知るための予備実験として，現実の Web ページを収集し各パラメータの実測を行った‡．その結果，$\bar{s} = 7.0 * 10^3$[Byte]，$\bar{t} = 3.3$[Sec] であった§．また，1Web サーバが持つ Web ページ数の平均 $\bar{p} = 2.6 * 10$ であった．

これより 10^7 程度の Web ページを収集する場合，N_{max} は，$3.8 * 10^5$ この際の $T_{wait} = 5$[Sec] とすると，$V = 4.6 * 10^4$[Page/Sec]，必要な回線容量は，$C = 4.6 * 10^4 * 7.0 * 10^3 * 8 = 2.5 * 10^9 bps$ となり，N_{max}，C 共に現実的な値とならない．従って，現実の情報収集ロボットの設計においては，N_{max} または C に利用可能な値を代入して，性能を設定することとなる．

3 実装と性能予測

3.1 情報収集ロボットの構成

[1] で報告した通り，我々が開発した情報収集ロボットは複数の PC を接続して協調して動作させることに

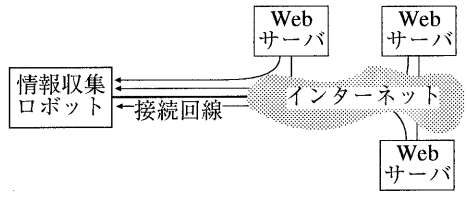

図 1: 情報収集ロボットの多重化

Gathering Speed Analysis for Internet Robots
Hiroshi TAKENO, Tokiharu NOTO
takeno@aether.hil.ntt.co.jp, noto@isl.ntt.co.jp
NTT CyberSolutions Laboratories
NTT CyberSpace Laboratories

‡ $\bar{p}$ は，無作為に抽出した URL $2.5 * 10^7$ より算出．他は実際に収集した Web ページ $7.7 * 10^6$ より算出

§ DNS キャッシュを使用しているため，同一 Web サーバへの連続収集が発生する場合 $\bar{t}$ は，単独で収集する場合より小さい値を取る

より同時接続数 N を確保するものである (図 2).

図 3 に，開発した情報収集ロボットのプロセス内構成を示す．同一の計算機上でこのプロセスを複数個動作させることが可能である.「収集 Thread」は N_p 個存在し，ソケットを通じて通信を行い結果を「受信データ」と「オフセット DB」に格納する.「既読管理 DB」,「統計情報」,「ドメイン情報」は，収集の管理，検索の補助に使うためのデータベースで，いずれも不可欠なものである．全ての収集スレッドは，これら 5 つのデータベースを独自に保有している．N_p は，コンパイル時に指定する．

3.2 環境と性能予測

表 1 に今回用意した環境を示す．これを用いて性能の推定を行う．

Solaris2.6 の，1 つのプロセスが同時に扱えるファイルの数の制限から $N_p = 150$ とした．1 台の PC に同時に起動可能な情報収集ロボットのプロセス数は，様々な要素の影響を受け算出することは困難であるが，プロセス走行中のメモリ消費量の実測値から 2 とした．これより，システム全体の同時接続数 N は，$150*2*5 = 1500$ となる．$T_{connect} = 5$ とすると $V = 180$ [Page/Sec] となり必要な回線容量は，$180 * 7.0 * 10^3 * 8 = 1.0 * 10^7$ [bps] となり，表 1 に示した回線容量の範囲に収まることが判る．

4 評価／考察

情報収集ロボットを実際に動作させた結果を 図 4 に示す．これより，次第に同時接続数が低下し収集速度に影響を与えていることが判る．全体的な収集速度は前章で推定した値の 40% 程度であった．これは収集が進むにつれて収集すべき Web ページの数が減少していることが原因であると考えられる．また，収集開始時ですら同時接続数が理論値の 60% 前後であることから，収集すべき Web ページの割り付けに問題があることが判る．収集すべき Web ページのアドレスの供給を安定させ，収集速度を維持することが今後の課題である．

参考文献

[1] 能登信晴，竹野浩，小橋喜嗣: インターネット検索サービスのための分散型情報収集，マルチメディア，分散，協調とモーバイルシンポジウム (DICOMO'98) (1998).

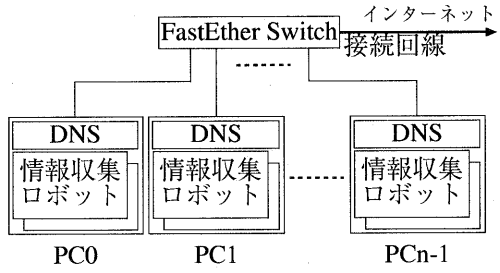

図 2: 分散情報収集ロボット

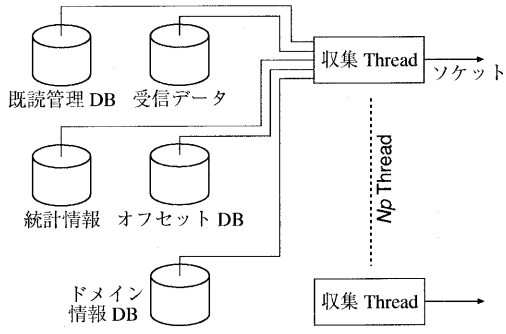

図 3: プロセス構成

表 1: 実験環境 と PC のスペック

接続回線容量	100Mbps
PC 数	5
CPU	PentiumuII 450MHz
RAM	640MB
HDD	18GB x 2
LocalNetwork	100Base-TX
OS	Solaris 2.6 (for Intel)

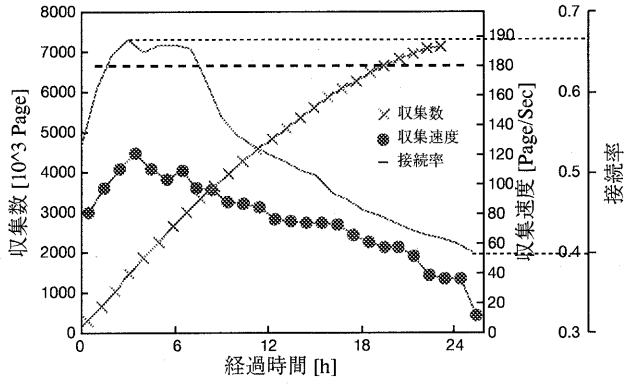

図 4: 収集特性