副テーマ報告書

分散型 WWW ロボットによる WWW 情報収集

学籍番号 0050078
氏名　松田　完
2001 年 2 月 20 日提出

１．はじめに

この論文では複数の WWW ロボットが、互いに重複しない WWW サーバのデータを強調して短時間に収集するための分散型 WWW ロボットを提案する。以下では、従来の WWW ロボットの問題を示すと共に、分散型 WWW ロボットを提案し、日本国内の WWW サーバを対象に 8 個所で分散収集を行っている状況について報告する。

２．従来の WWW ロボットの問題

WWW ロボットは、大きく分けて３つの問題を持っている。

まず、ネットワークと WWW サーバへの負荷の問題がある。これは一つの WWW サーバに対して検索システム毎にロボットが別々に動作しているためである。

次に、ページ数の増大による WWW ページ取得時間増大の問題が上げられる。WWW ページは指数関数的に増大しているのに対し、WWW ロボットを動作させるホストおよび回線容量がボトルネックとなっているためである。

最後に、最新情報の網羅性欠如問題である。WWW ページは頻繁に更新されるため、最新の情報に対する検索を実現するためには、WWW ロボットで再度収集して、検索システムの索引を作り直す必要がある。このため既に収集されたページについて更新チェックを頻繁に行わなくてはならない。

３．分散型 WWW ロボット

前節で挙げた問題点を解決するため、複数の WWW ロボットを強調動作させ、互いに重複しない WWW サーバのデータを収集することによりより、WWW データの収集を高速に行う。

分散型 WWW ロボットは図１に示すように全体を管理する Public Robot Server Manager(PRSM) と個々の WWW ロボットである Public Robot Server から構成される。収集されたデータは最終的に図中の Search Service Server に再配布することにより、検索サービスのためのインデックス作成を行う。

分散型 WWW ロボットにおいては、RPS と WWW サーバ間のデータ転送速度が、全体の収集時間を決定する大きな要因となる。このため、PRS の分担では以下に示すように、RPS と WWW サーバ間のサイト間距離を考慮した分散を行う。

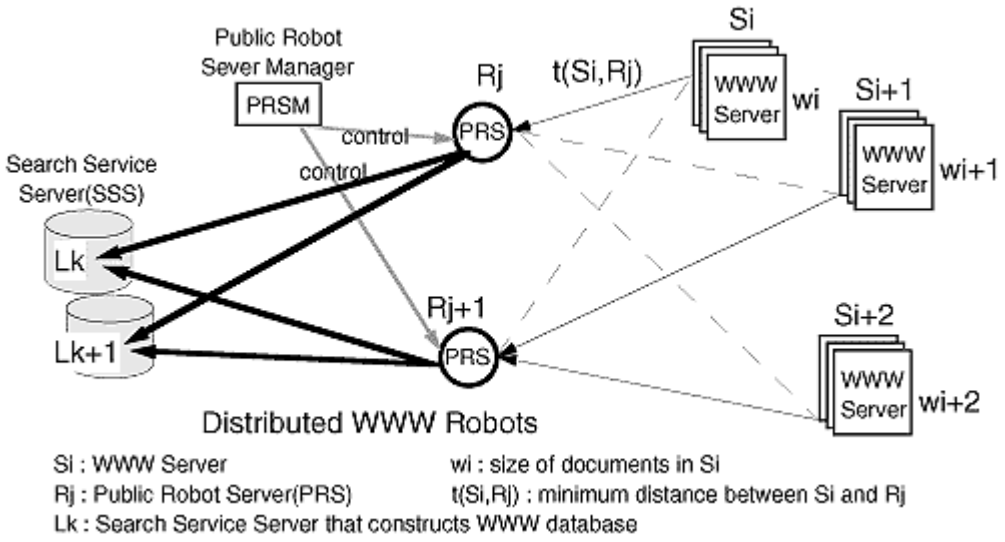

図 1: 分散型 WWW ロボットの定義

①モデルおよび定義

図 1 に示すモデルを用い $PRS(R_j)$ **が** WWW **サーバ**(S_i) **のデータを収集する収集コスト、及び、**$PRS(R_j)$ **が** Search Service Server(Lk) **へデータを再配布する再配布コストを定義する。これらのコストは、**S_i **のドキュメント量を** w_i **とすると、**w_i **とサイト間距離**$(t(S_i, R_j))$ **との積として定義できる。したがって、**WWW **サーバ** S_i **のドキュメントを** PRS R_j **が取得する場合、発生する収集コストは、**$w_i(S_i; R_j)$ **となる。また、**$PRSR_j$ **で収集した** WWW **サーバ** S_i **のデータを** Search Service Server Lk **へ再配布するコストは、再配布時のデータ圧縮率を** K **とすると、**$Kwit(R_j; Lk)$ **となる。**

②ネットワーク負荷と分散収集時間

全WWW **データ**$(S_i(i = ...n))$ **を**$PRS(R_j(j = 1...m))$ **で収集し、** Search Service Server(Lk(k = 1...l)) **へ全収集データを再配布する際に必要な収集コストと再配布コストの総和は、各**PRS R_j **の担当する**WWW **サーバ群を集合**V_j **で表すと**

$$\sum_{j=1}^{m} \sum_{S_i \in \{Vj\}} \{w_i \times t(S_i, R_j) + K \times w_i \times \sum_{k=1}^{l} t(R_i, L_k)\}$$

となる。これはネットワークに与える総負荷を示している。また、複数のPRS **の内、律速となる**PRS **が全体の実行時間を決定するため、分散収集時間は、以下の式で表すことができる。**

$$\max_{j=1,\ldots,m} \left(\sum_{S_i \in \{Vj\}} \{w_i \times t(S_i, R_j) + K \times w_i \times \sum_{k=1}^{l} t(R_i, L_k)\} \right)$$

③ネットワーク負荷および分散収集時間の最小化

分散収集時間を最小化するためにネットワーク負荷を最小にするアルゴリズムである欲張り法 (greedy

method) を用いた後、律速となる RPS への負荷を分散化することにより、分散収集時間を最小化する。

４．分散型 WWW ロボットを用いた WWW 情報収集実験

３節で述べた分散型 WWW ロボットを実装した実験結果を以下に示す。

4.1 サイト間距離 $t(S_i, R_j)$ の決定

サイト間距離 $t(S_i, R_j)$ として、ping が利用可能かどうかを調べるために以下の実験を行った。

4.1.1 ping 最小応答時間と HTTP 転送時間

ping の応答時間の最小値とHTTP 転送時間との間の相関関係を調べる実験を行った。3つのホストから3つの WWW サーバへに対し、1996 年11 月第 5 週の平日午前3 時頃に以下の測定を行った。

測定項目

- ping 最小応答時間(1032 byte パケット 10 個の最小値) から求めた転送速度
- HTTP により50 ページを転送した際の平均転送速度

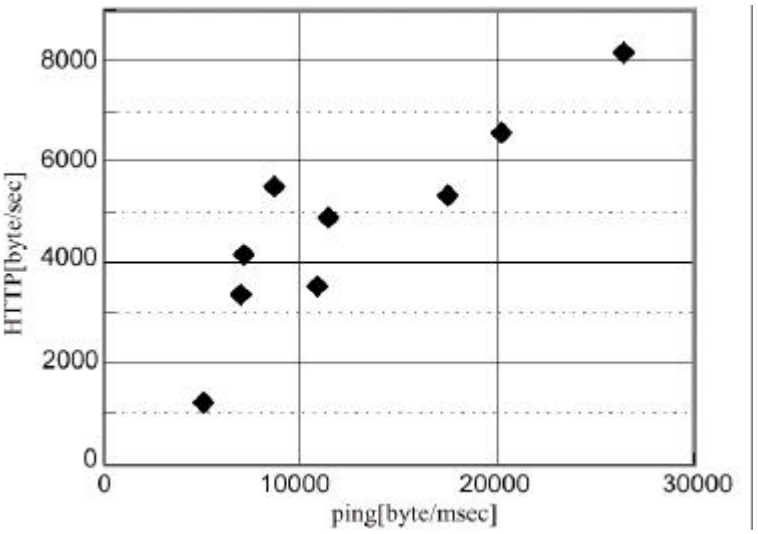

図 2: ping の最小応答時間から求めたデータ転送速度と HTTP 転送速度との相関関係

測定結果を図 3 に示す。ping 最小応答時間から求めたデータ転送速度と HTTP による平均転送速度の間に相関関係 (比例) があることがわかる。

4.1.2 HTTP 転送速度の予測

4.1.1 から、 ping 最小応答時間と HTTP 転送速度の間にほぼ比例関係が成立することがわかった。この関係を用いて転送速度を予測する手法について検討する。

1988 年 3 月 17 日午前 2 時〜午前 9 時にかけて電総研から 4 つの WWW サーバに対して以下の測定を行った。

測定項目

- ping **応答時間**(64/1024/2048 byte **パケット**10 **個の最小値**)
- HTTP **により**100 **ページを転送した際の転送時間**

表1**に測定結果を示す。**

表1: ping **最小応答時間**, Latency, Throughput

WWW Server	Ping **最小応答時間**(packet size[byte]			**往復** Latency	**最大** Throughput
	64	1024	1048		
	[ms]	[ms]	[ms]	[ms]	[byte/s]
(A)	20	40	55	17.7	54,949
(B)	18	45	61	15.2	44,716
(C)	29	57	74	26.1	42,729
(D)	75	103	117	72.3	45,773

表 2: HTTP **での転送ファイル数、平均サイズ、ドキュメント量、転送時間**

WWW Server	**ファイル数**	**平均サイズ** [byte]	**ドキュメント量**[byte]	**転送時間** [ms]
(A)	99	2811	278,259	21,677
(B)	76	3158	240,006	14,455
(C)	98	4526	443,501	31,491
(D)	97	2157	208,778	46,278

次に、4.1.1 **の結果より、転送時間が** ping **最小応答時間に比例すると仮定して、以下の式で近似し予測転送時間を算出する。**

$$\text{予測転送時間} = \frac{(ping\text{最小応答時間})}{(ping\text{パケットサイズ})} \times (\text{ドキュメント量}) \times (\text{補正パラメータ}\ a)$$

この結果、64byte **の** ping **の最小応答時間のみを用いて予測した場合でも実転送時間との誤差を** 14%**程度におさえられることがわかった。このため、サイト間距離として** ping **の最小応答時間を用いることにした。**

4.2 分散型 WWW ロボットによる実験状況

分散型ロボットにおけるインプリメントでは、PRS とWWW サーバ間のサイト距離の値として、ping を30 分おきに48 回行ないその最小値を$t(S_i ; R_j)$ として採用した。ドキュメント量(w_i) は事前には分からないため、初期値は1 とし、PRS が収集した総ドキュメント量が判明した段階で、定期的(現状では１ ヶ月毎) に再配置する仕様とした。また、第一ドメインがjp であるサイトを対象にデータ収集を行う。1997 年7 月より32 個のS_i (32 ドメイン) を初期値として設定し、早大、シャープ、東大、京大、ASCII、電総研、北陸先端大、慶大8 個所でPRS を動作させている。

1997 年1 月14 日の時点で、13320 ドメイン(第1 〜第3 レベルのドメインが同じドメインを１つのドメインと数える) を発見し、ping によるサイト間距離計測を継続中である。表3 に各PRS でサイト間距離計測が終了したドメイン数を示す。

Waseda U.	Sharp	Tokyo U.	Kyoto U.
1964	1456	520	347
ASCII	ETL	JAIST	Keio U.
298	2622	26	331

現時点では全てのドメインについてサイト間距離の計測が終了していない。しかし、既に終了しているものの中では 1.1〜10.9 倍、平均 2.8 倍の差が計測された。なお、S_i が RPS と同一の場合には 27 倍もの差が観測された。この結果から、各 RPS がネットワーク的に近いドメインを対象として分散収集すれば RPS の台数以上の高速化が得られること考えられる。4.1 節の結果を用いると、WWW ロボットを n 台に分散させた時、最大(2.8 × n)倍の高速化が得られると予想される。

5 終わりに

本論文では、分散型WWW ロボット実験の動作状況について報告した。当面の目標として日本国内のWWW サーバ上の全データを24 時間以内に収集可能な分散型ロボットの構築を目指して、実験を続けている。今後は、サイト間距離としてping を用いることの正当性を厳密に検証すると共に、サイト間距離として、前回のWWW ロボット巡回時のHTTP 転送時間の利用も考えていく。また、当面の目標である24 時間以内のデータ収集を達成を目指すと共に、次のステップとして、検索能力をPRS 内に収容した分散検索についても検討を行っていく予定である。