HORB を使用した分散 Web リンク収集システム

秋田 晋吾† 林 幸雄† 松久保 潤†
北陸先端科学技術大学院大学†

1.はじめに

近年のネットワーク環境の発展により、多くの計算機資源を効率よく利用する方法として、Grid コンピューティングをはじめとする広域分散コンピューティングの研究が盛んに行われている[1][2]。中でも、負荷均一化手法は重要な課題の一つである。

一方、ネットワークを透過的に扱い、記述の比較的容易な分散プログラミング環境も整備されてきた。

本研究では JAVA 用の分散オブジェクト技術 HORB[3] を使用して、分散協調型 Web リンク収集システムを構築した。このシステムを使用して Web リンク収集実験を行い、その性能や有用性について検討した。

2.分散 Web リンク収集システム

2.1.Web リンク収集タスク

Web リンク収集タスクとは URL から、そのリンク先にある HTML ファイルを採集し、ファイル内に含まれる新たな URL リンクから HTML ファイルを探索するという過程を繰り返すことにより、ネットワーク内に存在する URL を収集するタスクである(図 1)。

2.2.分散処理

本研究では複数の計算機でツリー状のネットワークを作り、ネットワーク全体で Web リンク収集タスクを実行する。その際、収集タスク、URL データの管理、計算機間の通信などをマルチスレッド処理で行う。

全ての計算機はネットワーク上で自分と隣接するノードとのみ通信を行う。ネットワーク内での通信は２種類あり、一つは各計算機が収集した URL の共有、もう一つは負荷を移動させることにより各計算機の負荷を均一化するものである。

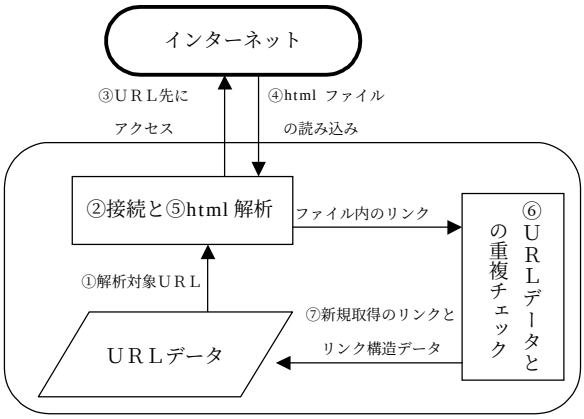

図 1.各サーバにおける基本処理

3.負荷均一化アルゴリズム

ネットワーク内の負荷を均一化するアルゴリズムは以下の通りである。

なお、ここではツリー状に構成されたネットワークの根にあたるサーバをルートサーバと呼ぶ。根を基準にして、各サーバには親子関係が定義できる。

1. 負荷量・稼働時間などをトリガーとして親サーバに負荷均一化を依頼する。依頼されたサーバがルートサーバでない場合、更に親サーバに依頼し、信号をルートサーバまで送る。
2. ルートサーバは自分の持つ子サーバに現在の負荷量を送信するよう要求する。依頼を受けたサーバは自身の子サーバの負荷量を要求することを繰り返し、ネットワーク全体の総負荷量を収集する。
3. ルートサーバから各サーバに均一な負荷量をブロードキャストする。
4. ネットワークの各サーバは親サーバと負荷移動を行う。自分の負荷量が均一負荷量より多い場合は親に負荷を送り、逆の場合は親の負荷を受けとる。

Distributed Web Link Collecting System with HORB
†Shingo Akita, Yukio Hayashi, Jun Matsukubo
Japan Advanced Institute of Science and Technology

4.収集実験

ネットワークは 10 台の計算機で構成し、接続は末端のサーバ以外が全て 3 台の計算機と接続されたツリー状(最も効率的なベーテ木)に配置した。また、本実験内での負荷量は、その計算機が持つ未探索 URL の数で定義した。

まずネットワーク内での負荷均一化を行わずに(10 台を独立して稼働させた)90 分間の Web リンク収集を行った結果を以下に示す。

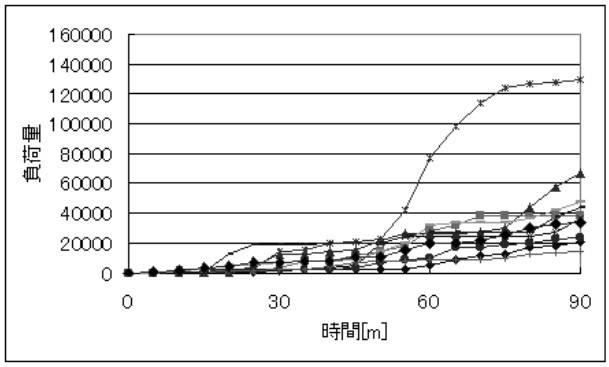

図 2.負荷均一化無しの負荷量の変動

各計算機の負荷量に差が生じ、負荷量が最大のものと最小のものでは 10 倍近い差が生じた。

次に、ネットワーク内で前述の均一化アルゴリズムを行った場合の負荷量の推移を示す。各計算機は５分ごとに自分の負荷量を記録し、前回の記録から負荷量が 10000 以上(高負荷の目安)増えている場合、均一化処理を行った。

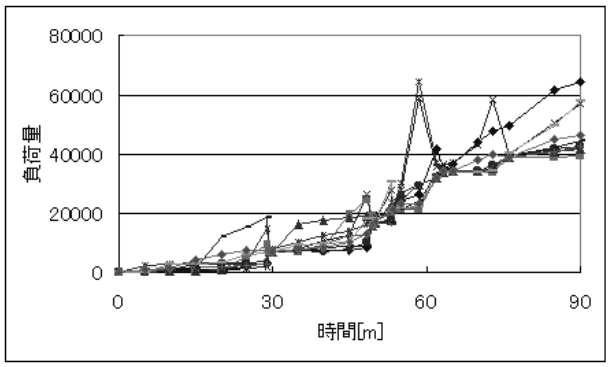

図 3.負荷均一化を行った実験の負荷量の変動

個々の負荷量は一時的に差が生じることはあっても、負荷均一化処理の効果がうかがえる。例えば、高負荷の発生していた 50 分過ぎのところで負荷均一化が行われ、10 台の計算機の負荷が均一となっている。

この二つのグラフの標準偏差を計算し、その推移を図 4.に示した。◆付き実線で示したグラフが大きく落ち込んでいる点が負荷均一化を行ったタイミングを表している。図 2.において高負荷の発生していた 50 分～70 分の時間帯に 3 回の負荷均一化を行っている。

また、どちらの実験においても計算機 1 台当たりの処理した URL 数は 3000～16000 となり、全体での処理 URL 数は約 83000 ページ、総獲得リンク数は約 550000 本であった。

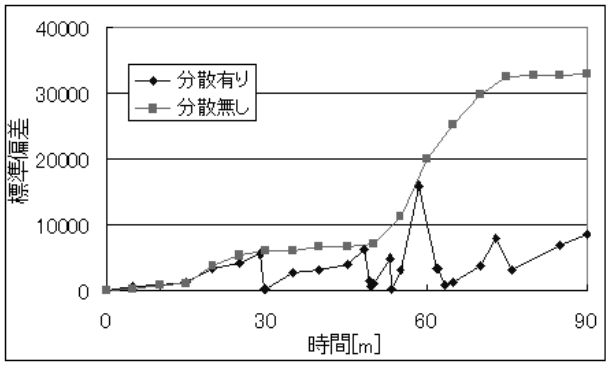

図 4.均一化処理の有無による標準偏差の違い

5.おわりに

本研究では、分散環境上での Web リンク収集システムを作成し、複数の計算機上で協調動作させることが出来た。また、ツリー状のネットワーク全体の負荷が均一化できることを示した。

また、本実験で使用した 10 台の計算機は全て同じ LAN 上に存在するものであったが、負荷の均一化開始から終了までにかかる時間は数秒～2 分以上と時間に大きく差があった。これらは回線の状態やデータ転送量に大きく影響されるものであるが、これが Grid の様な広域分散ネットワーク内では更に大きくなることが予想される。

参考文献

[1] 村岡洋一, ”Internet 広域分散協調サーチロボットの研究開発”研究成果報告書, 情報処理振興事業協会, 2000.

[2] 伊藤 正敬, 林 幸雄, “ORB 分散コンピューティングを用いた Web リンク収集”, 信学総合大会, D-6-17, Mar.27-30, 2002.

[3] http://www.horb.org/horb-j/