分散型 WWW ロボットの実験状況と今後の課題

山名早人
電子技術総合研究所　情報アーキテクチャ部
http://www.etl.go.jp/~yamana/DWR/

WWW ロボットは、HTML 中のリンクを自動的にたどることによって WWW サーバ上のデータを自動的に収集するソフトウェアであり、1999 年 10 月現在で約 211 種類が登録されている [1]。WWW ロボットによるデータ収集は、日本国内の WWW サーバのみを対象とした場合でも、WWW ロボットを動作させるサーバの性能および回線容量の制限などから、収集に 1 ヶ月以上の時間を要している。

本研究開発は、従来の WWW ロボットをネット上に分散配置し、協調収集することにより、ドメイン名が jp で終わる WWW サーバの全データを 24 時間以内に収集することを目標としている[1]。コアメンバとして、早大、京大、北陸先端大、慶應大、府立大、日本 IBM(株) 東京基礎研、シャープ (株)、電総研の 8 研究機関が参加すると共に、外部協力機関を併せて合計 32 機関が参加している。

分散型 WWW ロボットは、(1)WWW ロボットをネットワーク上に分散して複数配置し、(2) 各 WWW ロボットが担当する WWW サーバを全体の負荷が均一化されるようにスケジューリングすることにより、高速に WWW サーバ上のデータを収集する。これによって、(1) 検索サービスサイトでの最新データ検索の実現、(2)WWW ロボットがインターネットに与える負荷の軽減を目指す。

電総研の WWW サーバに対する各種 WWW ロボットからのアクセスは、全アクセス数の約 37 % (99.7.12-18 の平均) を占めている。さらに、http プロトコルが現在のインターネットの負荷の約 70 %を占めていることを考えると、WWW ロボットが個別にデータを収集せずに、協調収集することによるネットワーク負荷軽減の効果は大きい。

分散型 WWW ロボット (図 1) は、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(PRS) から構成される。PRS への担当 WWW サーバ割り当て方式として、ランダム方式と負荷均等化方式の 2 種類をサポートする [2]。収集されたデータは、Search Service Server(SSS) に再配布することにより、検索サービスのための索引作成を行う。

[1]情報処理振興事業協会 (IPA)- 独創的情報技術育成事業「インターネット広域分散協調サーチロボット研究開発」として研究開発が行われている

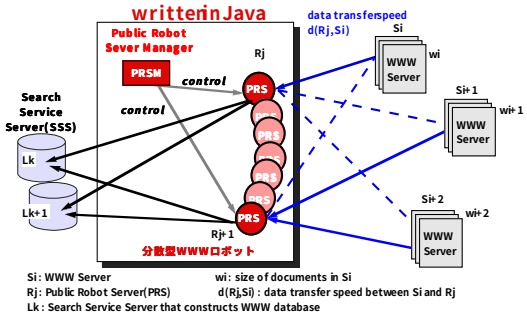

図 1: 分散型 WWW ロボット

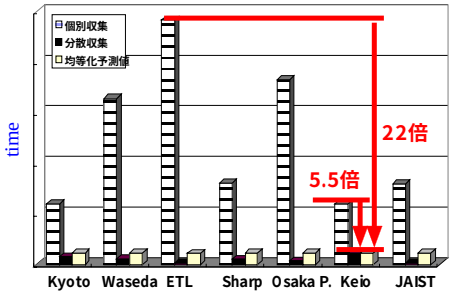

図 2: 分散型 WWW ロボットによる高速化 (7 分散)

日本国内の 103 箇所の WWW サーバを対象とした収集実験では、一カ所で集中して収集する場合に比較し、7 個所に分散することにより、ランダムな分散で 2.6〜10.6 倍の高速化が可能であり、負荷均一化を行った場合、5.5〜22 倍の高速化が可能であることがわかった (図 2)。

平成 11 年 11 月より、国内約 30 個所に分散型 WWW ロボットを配置し、本システムを実際に運用開始しており、今後、(1) 収集の効率化、(2) 負荷変動への対応の検討、(3) 広域分散環境でのメンテナンス性確保についてさらに検討を行っていく予定である。

参考文献

1) :「The Web Robots Database」, http://info.webcrawler.com/mak/projects/robots/active.html

2) -: 「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 (IPA) (1999.2)