Internet 広域分散サーチロボットの研究開発

Internet Wide-area Distributed Search Robots

村岡洋一 *1, 田村健人 *2, 山名早人 *3, 河野浩之 *4, 森 英雄 *5
浅井勇夫 *6, 西村英樹 *7, 楠本博之 *8, 篠田洋一 *9

*1 早稲田大学 理工学部
*2 日本アイ・ビー・エム 東京基礎研究所
*3 電子技術総合研究所　情報アーキテクチャ部
*4,*5 京都大学大学院 情報学研究科
*6 大阪府立大学 工学部
*7 シャープ (株) 技術本部
*8 慶應義塾大学 環境情報学部
*9 北陸先端科学技術大学院大学 情報科学研究科

本報告では、広域に分散した WWW サーバのデータの高速収集するための一手法として、インターネット上に分散した複数の WWW ロボットと呼ばれるプログラムを協調動作させる「分散型 WWW ロボット」について実験状況を示す。本ソフトウェアは次世代インターネットにおける「広域分散コンピューティング」分野における一つの重要なアプリケーションとなると共に、次のような利益をもたらすと考えられる。分散型 WWW ロボットを用いることにより、WWW サーバ上のデータを高速収集（目標は日本全国の WWW サーバ上のデータを 24 時間以内に収集) することが可能となると共に、検索サービス提供サイトでは、最新のデータを利用した検索サービスの提供が可能となる。これは、インターネット検索ビジネスの活発化にもつながる。これまでの実験結果より、分散しないこれまでの方式に比較して 7 つに分散した場合、5.5 倍〜 22 倍の性能向上が得られることが分かった。さらに、現在各検索サービスサイト毎に独立に動かしている WWW ロボットを本研究開発によって開発された広域分散協調サーチロボットに置き換えることにより、現在インターネットの負荷の 70 %を占める http プロトコルによるデータ転送の内、WWW ロボットによる負荷（約 25 %) を大幅に削減することが可能となる。

1　はじめに

World Wide Web (WWW) は、Mosaic が開発されて以来、インターネット上での情報提供及び情報収集の手段として世界中で利用されている。特に、情報検索サービスは、今やインターネットのポータルサイトとして認知されるに到っている。しかし、以下に述べるように多くの問題が残されている。

ブラウザが世の中に認知されはじめた 1993 〜 94 年頃には、検索サービスは存在しなかった。ところが、その後のいわゆる Mosaic や Netscape の登場に起因するインターネットブームにより、インターネット上で提供される情報が急速に増加した。これをきっかけとして、WWW の情報を検索するための仕組みの構築が始まり、現在の AltaVista [1] や HotBot [2] に代表される検索サービスへと至っている。

WWW サーバ数は、1999 年初めに全世界で 400 万台を突破し (図 1)、それらのサーバから発信される情報は 6 億を越えると推測される（97 年末で 3.2 億 [3] と推測されており、かつ、99 年初頭までに WWW

本研究は、情報処理振興事業協会「独創的情報技術育成事業」の一環として行われたものである。

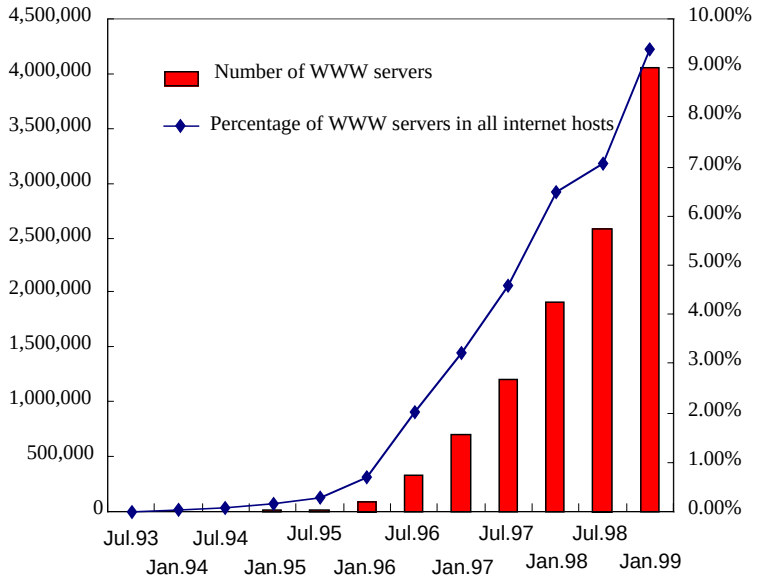


図 1: WWW サーバ台数と全ホスト数に占める割合の推移 ([4] [5] のデータに基づき作成)

サーバ数が倍増しているため)。さらに、現在もインターネットに接続されるコンピュータ台数の 2 倍以上の伸び率で WWW サーバは増え続けており、インターネット上で入手できる情報 (URL 数) は、たった 1 年で 2 倍以上に膨れ上がっている。

しかし、検索サービスで検索できる URL 数は、AltaVista（公表されている収集 URL 数では最大を誇る米国の古参の検索サービスサイト）でも 1.4 億 URL 程度であり、全体の 1/4 以下である。検索できる URL 数が全 URL に占める割合（カバー率）は、年々減少する方向にあり、「欲しい情報」は検索できるかもしれないが「検索されて出てきた結果が全てではない」という状況が発生している。

検索サービスサイトは、大きく Yahoo! [6] のようなディレクトリ型と、AltaVista や HotBot のようなロボット型に分類できる。ディレクトリ型検索サービスでは、WWW のアドレスを示す URL(Universal Resource Locator) を、人手により、芸術、ビジネス、教育..、のように分野別に分類する方式をとっており、データ量が少ない反面、人手で索引や要約を作成するため、索引と要約の信頼度が高いといった特徴を持つ。一方、ロボット型検索サービスでは、WWW ロボットやスパイダーと呼ばれる Web 探索プログラムを用いて、インターネット上で見つけることのできる WWW サーバー上の情報を定期的に収集し、その情報の索引付けを行っており、情報量が多いという利点を持つ。逆に、各ページの要約を自動的に生成したり、索引付けを自動で行うため、要約の完成度が低いという欠点を持っている。

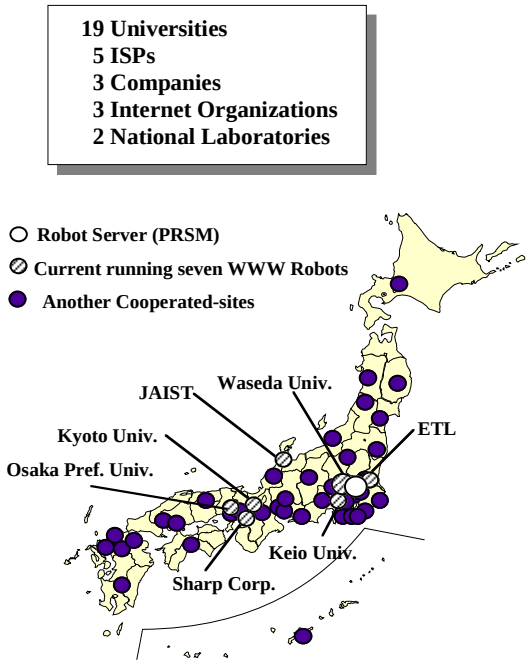
そして、ロボット型検索サービスを使っても、現在のインターネット上の急速な情報量増加に対応することができず、先に述べたように、検索できる URL 数が全 URL に占める割合が 1/4 以下になっている。

このような現状を打破するため、広域な範囲の WWW データをインターネット上に分散された複数の WWW ロボットで協調して高速収集するための「分散型 WWW ロボットの実験」を本研究で行っている。平成 11 年度中には、日本国内の全 WWW サーバ (ドメイン名が jp で終るサーバ) のデータを 24 時間で収集する仕組みを構築する予定である。以下では、これまでの実験状況を報告する。

2 研究開発目標と参加者

WWW ロボットが全ての WWW サーバにアクセスして必要な WWW データを収集するには、現状技術では、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、日本国内のみの WWW サーバを対象とした場合でも 1 ヶ月以上の時間を要している。

そこで、分散型 WWW ロボット（ソフトウェア）の研究開発では、日本国内 (ドメイン名が jp で終わる WWW サーバ) の全 WWW データを 24 時間以内に収集することを目標とした。



1. Univ. of Tokyo
2. National Center for Science Information System
3. JAIST(Japan Advanced Institute of Science of Technology)
4. NAIST(National Institute of Science of Technology)
5. Intitute of Industrial Science, Univ. of Tokyo
6. Gifu Univ.
7. AFFRC(Agriculture,Forestry and Fisheries Research Council
8. Sharp Corp.
9. Kyoto Univ.
10. Electrotecnical Laboratory
11. Keio Univ.
12. Kansai Univ.
13. Tokushima Univ.
14. Iwate Univ.
15. Tokyo Institute of Technology
16. Yokohama National Univ.
17. NTT Optical Netwok Systems Laboratory
18. DEODEO Internet Service Co.Ltd.
19. Ryukyu Univ.
20. Nanzan Univ.
21. Osaka Prefecture Univ.
22. Netwave Shikoku Co.Ltd.
23. TRY Network International Co.Ltd.
24. NORTH(Network Organization for Research and Technology in Hokkaido)
25. Tohoku Internet Association
26. NSK Internet Service Co.Ltd.
27. Waseda Univ.
28. IBM Tokyo Research Center
29. Intio Internet Service Co.Ltd.
30. Kyusyu Univ.
31. HIroshima Univ.
32. Tokai Internet Exchange

図 2: 実験参加組織

コアメンバとして、早稲田大学、京都大学、北陸先端大学院大学、慶應義塾大学、大阪府立大学、日本 IBM(株) 東京基礎研究所、シャープ (株)、電子技術総合研究所の 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（19 大学、5 インターネットサービスプロバイダ、3 企業、3 ネットワーク機関、2 国立研究所）が参加している (図 2)。

3 分散型 WWW ロボットの概要

研究開発を行っている分散型 WWW ロボットでは、

① WWW ロボットをネットワーク上に分散して複数配置する。

② 分散した WWW ロボットが担当するサーバを自動的に決定させると共に、協調動作させる。

の 2 点により、高速に WWW サーバ上のデータを収集することを目指しており、直接の効果・成果、及び、波及効果として以下の 4 点が期待される。

① WWW サーバ上のデータを高速収集（目標は日本全国の WWW サーバ上のデータを 24 時間以内に収集) することが可能となる。

② 高速収集により、検索サービスサイトでは、最新のデータを利用した検索サービスの提供が可能となる。つまり、最新性において質の高い検索を提供することができるようになり、インターネット検索ビジネスの活発化にもつながると予想される。

③ 現在各検索サービスサイト毎に独立に動かしている WWW ロボットを本研究開発によって開発された広域分散協調サーチロボットに置き換えることにより、国レベル、あるいは、世界レベルでの協調収集が可能となる。これにより、現在インターネットの負荷の 70 %を占める http プロトコルによるデータ転送の内、WWW ロボットによる負荷を大幅に削減することができる。電子技術総合研究所の WWW サーバに対する平成 11 年 7 月 12 日〜 18 日のアクセスを調査したところ、全アクセスの 37 %が WWW ロボット

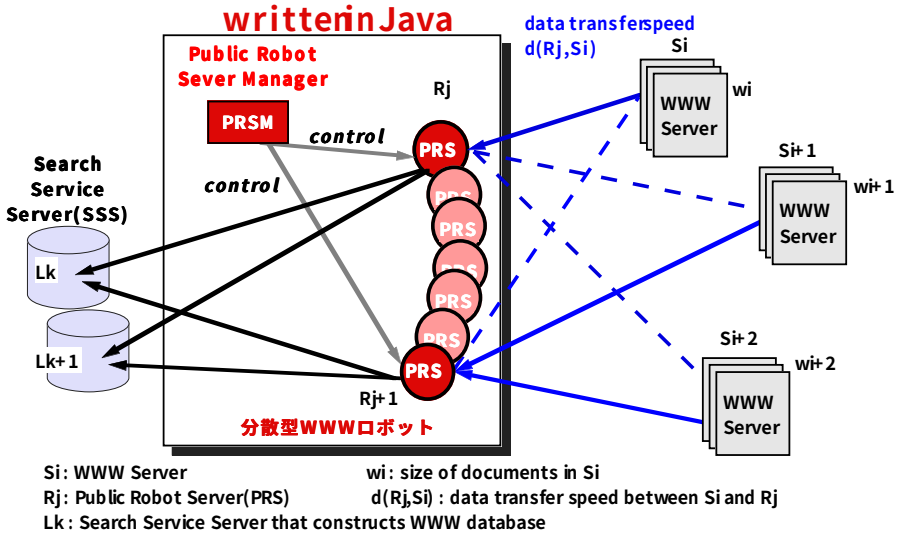


図 3: 分散型 WWW ロボットの仕組み

によるものであることがわかった。すなわち、インターネットの負荷の内、70 %×37 %＝約 25 %が WWW ロボットによるものであり、この 25 %の負荷を大幅に削減することが可能となる。

④ 本ソフトウェアは次世代のコンピューティング環境として近年研究開発が活発化の兆しを見せている「広域分散コンピューティング」の一つの重要なアプリケーションとなり、本分野の研究開発の活性化に寄与する。

3.1 分散型 WWW ロボットの動作概要

分散型 WWW ロボットは、図 3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(PRS) から構成される。

PRSM は、PRS に対して担当 WWW サーバの分配や、各 WWW サーバと PRS 間との距離計測を指示する。一方、PRS で新規に発見された WWW サーバや距離計測の結果は、PRSM に送られ、PRSM 側で PRS への担当 WWW サーバ割り当てを行う。PRS への担当 WWW サーバ割り当て方式としては、現在、ランダム方式と負荷均等化方式の二種類をサポートしている [7]。このようにして、PRS は PRSM からの指示に基づき各々互いに重複しない WWW サーバを担当し WWW データを収集する。

収集されたデータは、最終的に図中の 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。現在の HTTP/1.1 準拠のサーバは、1 回の接続で複数のデータを転送する Keep-Alive 機能を持っているが、複数のデータをまとめて転送する機能は持っていない。このため、各 PRS で収集したデータを SSS に再配布する際のオーバヘッドを小さくするため、複数のデータを 1 回の接続でまとめて圧縮して送る機能、及び、SSS からの要求に応じてデータ中の必要な部分のみ (例:URL や更新日時の指定による指定) を送る機能を PRS に持たせる。

現在、PRS・PRSM は Java1.2 で実装されており、PRSM は WWW サーバ Apache に JServ モジュールを組み込み、処理は Java servlets で行っている。

3.2 PRS の動作

PRS は、まず起動時に PRSM にアクセスし、担当リストを PRSM から取得する。担当のサーバのページをすべて取得し終ると、再び担当リストを取得する。

ページ収集は、インターネット及び相手の WWW サーバへ与える負荷を最小限におさえるため、動作時間を制限し、平日は午前 2 時～午前 8 時、土日は午前 2 時～午後 10 時のみページ収集を行う。これは、Internet Exchange サイトの調査で、午前 2 時～午前 8 時頃の負荷が一番小さいとの結果に基づいている。さらに、Java のスレッドを用い、同時に 200 個のサーバに対して並列にアクセスを行ない、高速化を実現し

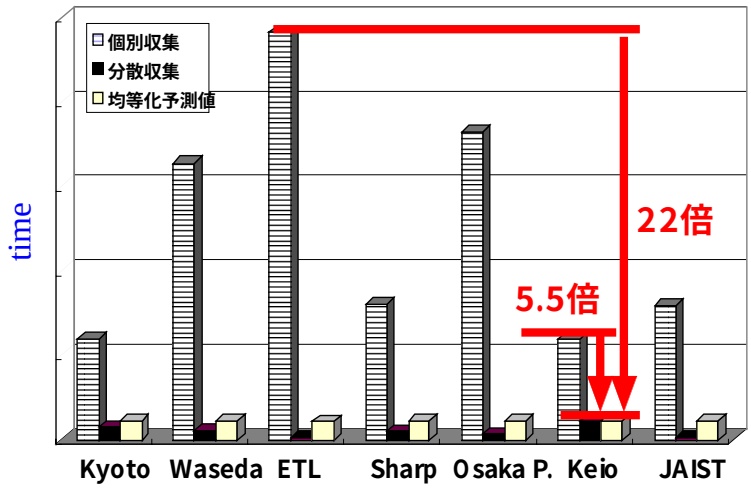


図 4: 分散型 WWW ロボットによるデータ収集の高速化 (7 分散時)

ている。

1 つのサーバに対しては、サーバへの負荷を少なくするため、連続してアクセスはせず 20 秒おきにアクセスを行なうが、WWW サーバが HTTP/1.1 の Keep Alive 機能に対応している場合は可能な限り連続でアクセスを行なう仕様となっている。

なお、相手の WWW サーバの能力や負荷状況に応じてアクセス間隔を変更できる仕組みについては、今後の課題である。

4 これまでの実験結果

平成 10 年度作成したプロトタイプシステムを用い、日本国内の 103 の WWW サーバを対象とした収集実験を行い、実際の分散型 WWW ロボットの有効性を確認した。具体的には、一カ所で集中して収集する場合に比較し、7 個所（早大, 京大, 北陸先端大, 慶應, 府立大, シャープ, 電総研）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 4)。

4.1 負荷均等化方式

負荷均等化方式では、担当する WWW サーバが持つ「データ量」と「通信遅延 (レイテンシ)」の積を仕事量として定義し、各分散型 WWW ロボット (PRS) の仕事量が同じようになるように分散している [7]。これによって、ランダムに分散する場合に比較して、「負荷均等化分散」により、約 2 倍の性能向上が得られることがわかった。

一方で、通信遅延（レイテンシ）は、どの PRS が担当するかによって大きく変動し、図 5に示すように、同一の WWW サーバに対する各 PRS からの通信遅延は、平均 2.8 倍、最大で 10 倍以上の開きがあることがわかった。このため、うまく分散することにより、ｎ台の PRS を使った場合でもｎ倍以上の速度向上が得られることが判明した。平均が 2.8 であるので、最大ｎ × 2.8 倍の速度向上が期待できる。

4.2 収集時間変動

インターネットは、その構成が日々刻々と変化すると共に、ネットワークやサーバの負荷も曜日、時間帯によって大きく変動する。このため、本実験でも、計算上負荷均等化を実現しても、実測値では、計算上の予測値との間にズレが生じた。

図 6は、平成 11 年 2 月 11 日 (祝) の午前 2 時 ～ 8 時の間に収集した実測データを元に負荷均等化計算を行い、平成 11 年 2 月 16 日 (平日) の午前 2 時～ 8 時の間に、負荷均等化方式を用いた分散により実測したデータである。図を見ると明らかなように、実測値と計算上の値には大きな差が生じている。

面白いことに、実測値は、すべての場合において、予測値よりも小さな値を示している。これは、予測の基準となった日は、休日の夜であり、インターネット利用の負荷が上がり、平日時に比べれば収集に時間がかかったためだと推測される。そして、この結果、全体的に均等分散時の収集時間 (実測値) が短くなったと考えるのが妥当である。実際に、府立大に設置された

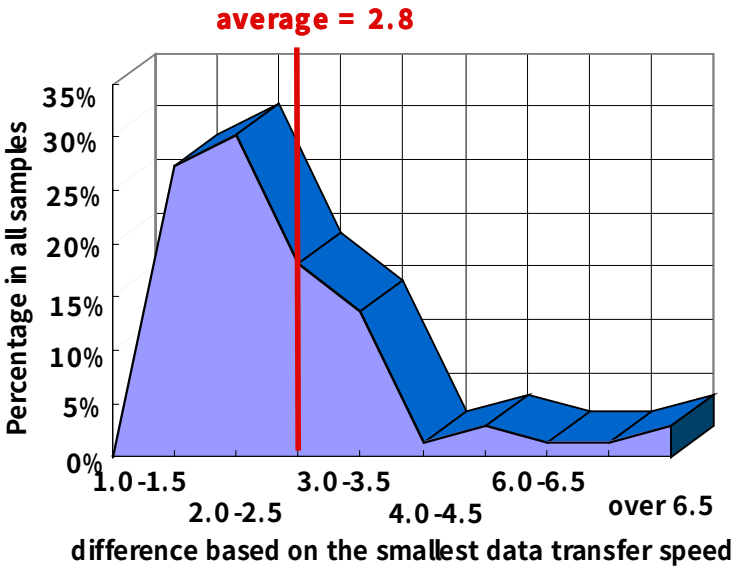


図 5: データ通信遅延 (レイテンシ) の分散

PRS のデータを解析したところ、URL 数が異なるので正確には判断できないが、平均 2.9 倍の収集時間の差があることが確認できた。

4.3 WWW サーバの能力に起因する収集時間変動

図 7は、さらに別の日についても実測を行い、予測値とのズレを予測値を基準としてパーセンテージで示したものである。予測の基準となるデータはを同じく平成 11 年 2 月 11 日 (祝) 午前 2 ～ 8 時に収集したものである。一方、均等分散時のデータは、同年 2 月 16 日と 17 日の午前 2 時～ 8 時に収集している。

図より、シャープ及び北陸先端では、16 日と 17 日の同じ平日でも実測値と予測値のハズレ方が大きく異なっている。これに対して、他のサイトでは、ランダム分散時と均等分散時の予測値との誤差にあまり違いがない。

この原因をログを解析することにより調べた。すると、シャープのランダム収集時の例では、***.ac.jp からの収集時間が予測の 10 倍の時間となっており、この特定サイトの収集時間だけでシャープにおける収集時間の 71% を占めていた。さらに、北陸先端の例でも、ランダム収集時に、特定の ***.co.jp からの収集時間が予測の 14 倍の時間となっており、この特定サイトの収集時間だけで北陸先端の収集時間の 19% を占めていた。

そこで、このような、予測と 10 倍以上異なっている特定サイトの収集時間が予測通りだったと仮定して予測値からの誤差を再計算した表を図 7(2) に示した。すると、予測と 10 倍以上の差があるサイトを除けば、予測値と実測値との間で、同じ平日に収集した場合の誤差がほぼ同じになることがわかった。

4.4 実験の結果判明した点

これまでの議論から、「収集時間は曜日による変動が大きい」、すなわち、土日や祝日等の午前 2 時～ 8 時は、平日の午前 2 時～ 8 時に比較してネットワークや WWW サーバの負荷が高いことがわかる。また、同じ平日でも、ほぼ収集時間が同じになる WWW サーバと、10 倍以上も収集時間が異なる WWW サーバが存在することがわかった。

つまり、能力の低い WWW サーバに対しては、収集時間の予測が困難であり、予測時に「能力の高い WWW サーバ」と「低い WWW サーバ」とを分類して、負荷均等化方式を適用しなければならないということがわかる。

5 収集データの再配布システム

分散収集された WWW データは、実験参加者間で共有することとしており、分散収集されたデータの再配布を行うシステムについて基本設計を行った。以下

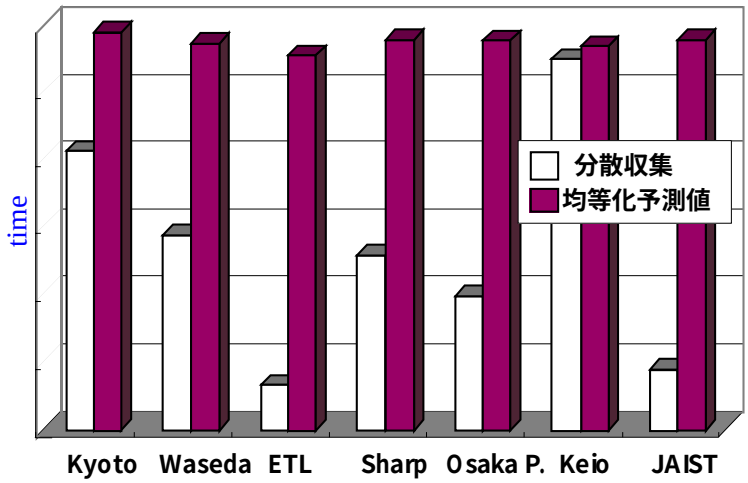


図 6: 収集時間の予測値と実測値の違い

にその概要を示す。なお、著作権法上の問題から、収集に参加していない第三者への配布は行わない。

全国 30 数サイトに配置されている PRS マシン上の指定ディレクトリ上にあるファイル (WWW データ) を再配布サーバ側に一日一回集め、実験参加者に再配布ができる仕組みを構築する。

- PRS →再配布サーバへのデータ転送

  PRS から再配布サーバへは、一日一回、サーバ側からの要求に基づいて差分データを転送する。転送時にはデータを tar+gzip 形式で圧縮して送る。セキュリティ確保のため、指定される再配布サーバ以外への転送はできない仕様とする。

- 再配布サーバでの処理

  各 PRS からのデータ転送スケジュールをファイルにより指定し、そのスケジュールに基づいてデータを PRS より転送する。サーバ側では、PRS 毎にデータ転送履歴を保持し、その転送履歴を元に各 PRS に対して差分転送を指示する。また、転送に失敗した場合には、1 時間後にリトライを行い、再度失敗した場合には、転送スケジュールにより指定される次の日に処理する。

- 再配布サーバからの再配布処理

  再配布サーバでは、各日の差分を合計 1 週間分保持し、再配布の効率化を図る。再配布サーバは、転送要求 (ホスト名指定、更新期間指定) に基づいて、 WWW データを圧縮した形式で再配布する。また、セキュリティを保持するため、あらかじめ登録されているサイト (IP アドレス登録) へのみ認証により再配布を可能とする。

6 今後の展開

平成 11 年 8 月までに、 3 サイト (早大, 京大, 電総研) で全 WWW サーバ (約 5 万サイト) を対象に試験的に 50URL/ サイト分の収集を実施した。これによって、 PRS 及び PRSM のバグフィックスを完了した。

平成 11 年 9 月からは、バグフィックス版の PRS の配布を開始し、平成 11 年 10 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する予定である。また、 PRS から再配布サーバへのデータ収集については、平成 11 年 11 月から開始することとしている。

さらに、平成 11 年 10 月からの実運用においては、以下の検討を詳細に行っていく予定としている。

- 負荷変動への対応の検討

  収集にかかる時間の曜日別格差がどの程度あるかを試験的に調べ、負荷均等分散の基準データを取得する。さらに、同じ平日でも、収集時間に 10 倍程度以上差がでる WWW サーバとほとんど収集時間に変動がない WWW サーバを分類できるかどうかについて検討し、均一分散を行う時に考慮する必要があるかを調べる。

- 広域分散環境でのメンテナンス性確保

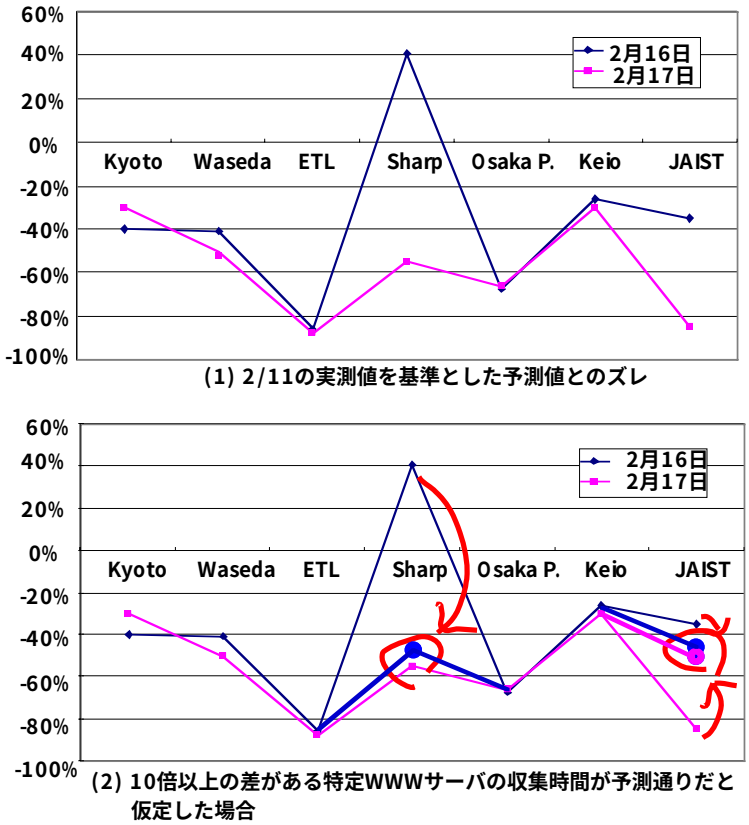


図 7: 1999 年 2 月 11 日を基準とした収集時間の曜日変動

広域に分散した複数のコンピュータを使った実験を行う際には、システムのメンテナンス性を高めることが実験効率化の必須条件である。このため、PRS のソフトウェアを自動的にバージョンアップするシステム等について、今後検討を行う。

7 おわりに

本報告では、広域に分散した WWW サーバのデータの高速収集するための一手法として、インターネット上に分散した複数の WWW ロボットと呼ばれるプログラムを協調動作させる「分散型 WWW ロボット」について実験状況を示した。分散型 WWW ロボットを用いることにより、WWW サーバ上のデータを高速収集（目標は日本全国の WWW サーバ上のデータを 24 時間以内に収集) することが可能となる。さらに、現在各検索サービスサイト毎に独立に動かしている WWW ロボットを本研究開発によって開発された広域分散協調サーチロボットに置き換えることにより、現在インターネットの負荷の約 25 %を占める WWW ロボットによる負荷を大幅に削減することが可能となる。

参考文献

1) AltaVista, `http://www.altavista.com/`

2) HotBot, `http://www.hotbot.com/`

3) Steve Lawrence, C. Lee Gilesd: *Searching the World Wide Web*,Vol,280, No.5360, Issue 3, pp. 98 - 100 (1998.4)

4) Network Wizards Internet Domain Survey: `http://www.nw.com/zone/WWW/top.html`

5) Netcraft Web Server Survey: `http://www.netcraft.co.uk/survey/`

6) Yahoo!, `http://www.yahoo.com/`

7) 「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)