

# ORB・分散コンピューティングを用いた Web リンク収集

## Gathering Web Links by ORB Distributed Computing

伊藤正敬

林幸雄

Masanori Itou

Yukio Hayashi

北陸先端科学技術大学院大学 知識科学研究科

School of Knowledge Science, Japan Advanced Institute of Science and Technology

### 1. はじめに

近年、増大する Web 情報から有用な Web ページを探すために、Web のリンク構造に着目した研究や検索サービスが盛んになっている。PC やネットワークの高速化など急速な IT の進歩により、複数台の PC で構成される比較的小規模な分散環境下でも、膨大な Web 情報を収集・解析することが可能になりつつある。

本研究では分散オブジェクト技術 HORB を用いた分散コンピューティング環境下で、Web のハイパーリンク収集システムを構築した。分散環境では Master - Slave 方式の負荷配分を採用しており、Master は初期条件（解析する URL 数、リンクの深さ）に基づいた Slave のタスクの割り当て管理を行う。

本論文では、まずシステム構成の概要、全体の処理フローについて述べる。次に、リンクを収集する SlavePC の台数変化による収集結果を示し、台数変化や初期条件変化による収集パフォーマンスに現れる影響についても述べる。

### 2. システム構成の概要

#### (1) 使用したマシン構成について

マシンは Pentium2-400MHz, 128M-SDRAM で、MasterPC 1 台、SlavePC 10 台、PostgreSQL データベースサーバ 1 台で構成した。

#### (2) 分散オブジェクト技術：HORB について

分散オブジェクトの代表的な技術には JavaAPI の RMI, CORBA, HORB があるが、HORB は IDL レス, SunJava 完全互換, RMI 以上の速度パフォーマンスを持っていることから、マルチスレッド処理を含めて容易かつ高速処理を実現できる技術と考え、本システムの実装に取り入れた。

### 3. システム全体の処理フロー

Master は Slave のタスク量を定める URL 数とリンク収集の深さの初期変数を持ち、データベースから取得した URL データを Slave に渡す。Slave は割り当てられた URL 先の Web ページを解析しリンクを取得する。また Master は常に各 Slave のタスク処理の監視をしており、Master はタスクを終了した Slave に新しい URL タスクを渡す。

また、Slave はタスクを終了すると、取得したリンクデータを Master に渡す。Master は渡されたリンクデータをデータベースに保存する。また Master はリンクデータから未解析 URL を新たなタスク URL として保存する。

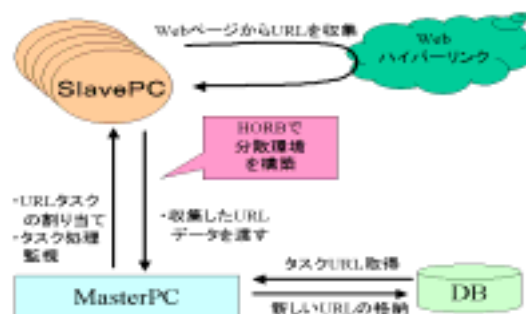


図1. システム全体の処理フロー

### 4. 台数変化によるリンク収集の実験例

例として Slave の台数変化 (2, 5, 10 台) によるリンク収集の結果を表1に示す。本実験では初期 URL 数3, リンクの深さ4とした。初期 URL は無作為に抽出した各大学の TOP ページとした。評価は「解析した Web ページ数」と「取得したリンク数」で行った。

|      | 2 台   | 5 台   | 10 台  |
|------|-------|-------|-------|
| ページ数 | 1586  | 3384  | 6033  |
| リンク数 | 20010 | 27227 | 43043 |

表1. 台数変化による URL 取得結果

### 5. まとめ

表1から、SlavePC を増やすと解析したページ数、取得したリンク数はともに増加していることがわかる。今回は初期条件を変えた結果を載せていないが、初期 URL 数、リンクの深さを大きくすれば取得できるリンクデータは多くなると考えられる。但し、時間などの他の評価を取り入れるとパフォーマンスは変わると考えられる。また初期に渡す URL データによっても結果が変化することも考えられる。URL の割り当て方や、リンク取得の探索方法などを考慮すればより効果的に収集が可能になるだろう。