静岡理工科大学　理工学部　情報システム学科　卒業論文

「並列分散型 Web データ収集システムの改良」

0513049　黒田博雅　［担当教員：幸谷智紀］

2009 年 2 月 16 日

1. 研究目的

本研究の目的は竹口友大先輩が開発した Web データ収集システム（Web Robot）を改良することである。このシステムでは収集した Web ページからのリンク URL を集めた順に次回の収集対象としているため、次の問題があった。

・同じドメイン内へのリンクが多い大手サイトからは出にくく（深さ優先)、収集内容が偏る。

・Web Robot を並列に動作させると、同じサーバへ同時にアクセスする可能性がある（アクセス攻撃と見なされる)。

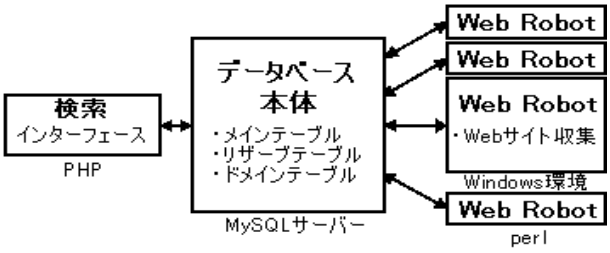
これらの問題を解決するため、次のような改良を行った。

・ドメインテーブルを作りドメイン別の収集を行う。

・保有 URL が少ないドメインを、URL が短い順に処理し浅く広く収集する。

2. Web データ収集システムの改良

以下の図はシステム全体の概要である。



Web Robot が Web サイトにアクセスし本文データをデータベースに保存する。検索インターフェースにキーワードを入力して検索開始、データベースに問い合わせて結果を表示する。

この Web Robot とデータベースに対して改良を行い、次のような収集手順に変更した。

手順１：担当ドメイン問い合わせ、保有 URL が少ないドメインを割り振る。(収集サイトの増加)

手順２:ドメインに該当する URL を”長さの短い順”に予約 URL(@url)に入れる。(浅い階層に外部へのリンクが多いので収集初期にドメインを増やす)

手順３：予約 URL にアクセスし収集する。

手順４：貼られている URL を保存、長さの計算とドメインの情報の更新を行う。

手順５：担当 PC を収集済みにチェック。(-1 に書き換える。周回カウントは全て-1 になったらリセット)

3. ベンチマークテストと考察

改良したシステムの性能を次に示す環境で計測した。

使用 PC：デスクトップ、並列分散収集に必要

CPU：Intel Core2 Quad 6600

メモリ：4GB

使用ソフトウェア

OS：　Cent OS 5.2 x86_64

Apache：Ver 2.2.3

MySQL：Ver 5.0.45

Active Perl：Ver 5.8.8

実験条件：1 日間、Yahoo Japan のトップから開始、フレッツ光の OCN 回線を使用した。

結果はこのようになった。

	ドメイン数	収集済 URL 数	予約 URL 数
旧型	1,992	3,072	113,587
新型	13,614	4,887	174,398

旧型に比べて約 7 倍近いドメイン数を取得することが出来た。収集した Web サイトの多様さの増加に加えて全体的に性能が向上したといえる。

次に分散処理の性能を計測した。

分散	ドメイン数	収集済 URL 数	予約 URL 数
単独	13,614	4,887	174,398
2 分散	29,000	5,001	240,873
4 分散	34,916	4,065	259,189

単独を 2 分散（2 プログラム）と 4 分散の両方とも上回ったが､2 分散と 4 分散の違いが余り出なかった。ドメインを該当 URL が少ない順に処理するため処理速度が上がり、ドメインテーブル、及びドメインを元に参照されるリザーブテーブルにアクセスが重なったことが速度の低下の原因になったと推測される。