2008 年 2 月 16 日

平成 20 年度　静岡理工科大学　理工学部　情報システム学科　卒業論文				
テーマ	Web データ収集構造を視覚化するためのプログラムの作成			
学籍番号	0513134	氏名	吉田　嗣	幸谷研究室

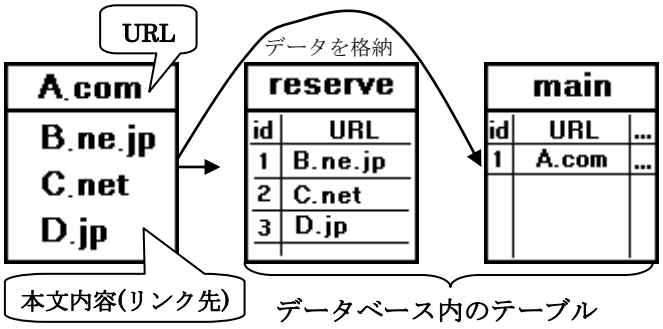
１．研究目的

本研究の目的は、データ収集プログラムの収集構造を、他人に説明し易くするためのプログラムの作成である。そのプログラムというのは、当研究室で過去二年間、研究開発してきたサーチエンジンの一部であり、それを用いて本研究を進めてきた。

本研究で作成したプログラムは、集められてきたデータをドメイン毎にし、図にして視覚化させ、説明の手助けをするものである。

２．システム概要

データ収集プログラムは以下の構造通りである。



データベース内のテーブル

初期 URL を設定し、その URL 内のリンク先を全て SQL 内の reserve テーブルへ格納。その URL のページ情報を main テーブルへ格納する。以降は、reserve テーブルに格納されている URL を参考に手順を繰り返し、reserve テーブルからその URL を削除するのが、プログラムの構造である。

次に、私自身が行った研究の概要の説明をする。

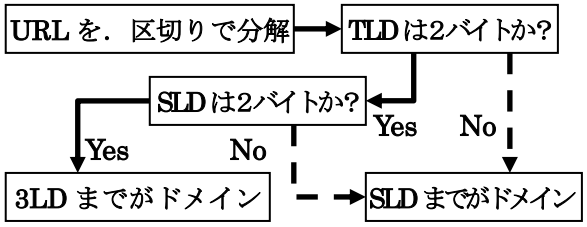
まずは、main テーブルに格納されている URL,refer(どの URL からきたものか)を読み込み、id に対応させて格納する。URL と refer をドメインに分解させ、ソートした後に、domains テーブルへ格納。その domains テーブルのデータを元に、ツリー状の図にして表示する。

３．ドメインと階層判定

ドメインというのは、URL の「 http:// 」から、次の「/」までの文字列のことである。仮に、URL を「http://yahoo.co.jp/」とした場合、ドメインは「 yahoo.co.jp 」いうことになる。右から、「.」区切りで、**トップレベルドメイン(TLD)**、**セカンドレベルドメイン(SLD)**、**サードレベルドメイン(3LD)**、…という。

yahoo.co.jp の場合、jp が TLD で、co が SLD、yahoo が 3LD となる。

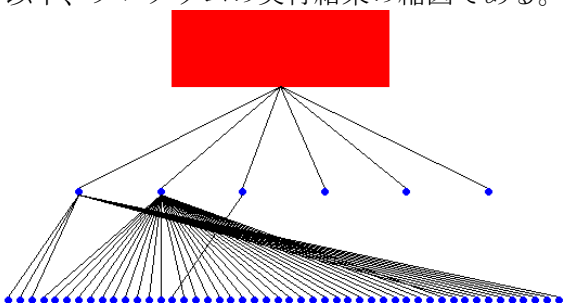
日本では、「**.jp**」という TLD が使用されている。個人サイトでは任意の SLD が取得可能であるが、組織・団体の場合、２バイトの SLD が割り当てられ、3LD の取得となる。国外では、そうなっておらず、２バイト以上の組織･団体用の SLD が存在するので、やむを得ず、以下のように判定することにした。



次に、階層構造の説明である。収集されてきた web データの構造に準じるため、重複するドメインは、単に一緒にせず、refer と階層が一致した時のみ統合させる仕組みである。初期 URL の階層を１とし、その URL を refer とする URL には＋１の値を足し、階層を判定していく。例えば、以下の場合は、B,C は A からきたので２ 、D は B からなので３となる。

４．結果

以下、プログラムの実行結果の縮図である。



まだソート方法が甘く、線が交差してしまっており、不本意ながら満足のいく段階まで到達しなかったが、樹形図の線画には成功した。階層・refer の同じものまとめて並べる課題が残ってしまった。

もし、この研究を引き継ぐ後輩がいたら、是非ともこの問題を解決していただきたい。