Web ページ自動取得システムによるデータの収集と解析

静岡大学 工学部 システム工学科 4 年
若林研究室 5061-3041 鈴木秀一

1 はじめに

1994 年頃からインターネットは急速に発達し，WWW 上で提供されている HTML のページ数は莫大な量になっている．膨大な情報の中から，必要な情報を瞬時に，かつ，的確に見つけ出すためのシステムが WWW 情報検索エンジンである．

現在の WWW 検索エンジンのひとつに，Web ページを自動的に回収する WWW ロボットを用いるロボット型検索エンジンがある．ロボット型検索エンジンは，Web ページ単位で検索を行い，その Web ページの内容による検索しか行えない．しかし，サイト (同一ドメイン内の Web ページの集まり) の作者が同一である場合が多いため，サイト単位で内容などの様々な特徴が表れる．一方，Web ページには，そのページの内容以外にもデータの種類などの様々な特徴があり，これらの特徴も検索に用いることができれば便利である．

本研究では，WWW ロボットを試作して，サイトごとに Web ページを収集し，データの種類や量と HTML ファイルのサポート状況の情報を提供するシステムの開発を行う．

2 システムの実装

実装したシステムは与えられた URI のドメイン以下の Web ページをすべて取得し，データを解析して，ドメイン内の統計的なデータを表示するものである．このシステムは，Web ページ自動取得部分とデータ解析部分で構成した．

2.1 Web ページ自動取得

Web ページ自動取得部分の主な機能は，ページ取得機能とリンク先取得機能である．

- Web ページ取得
  URI リストの URI から Web ページの情報を取得し，その情報から WWW サーバに接続し，Web ページを取得，保存する．
- リンク先取得
  取得した HTML ファイルからリンク先が指定してあるタグを識別して，そのリンク先の URI を取得する．これを補完して，同一ドメイン内の URI のみを URI リストに追加する．

2.2 データ解析

データ解析部分の主な機能は，ページの種類と大きさの情報を抽出すること及び，HTML ファイルのサポート状況をチェックする機能である．

- 情報の種類や大きさ
  Header 情報に含まれる Content-Type(データの種類) と Content-Length(データの大きさ) を，取得した Web ページ全てについて調べる．そして，全データ量，ページ数，1 ページあたりのデータ量，データの種類ごとのデータ量と割合を計算し表示する．
- HTML ファイルのサポート状況のチェック
  様々な規格またはブラウザでサポートされているタグと属性の一覧から，取得した HTML ファイルのタグの中でサポートされていないタグの個数を計算する．そして，全タグの個数とサポートされていないタグの個数から，サイトの中の HTML ファイルが様々な規格にどの程度 対応しているかを表示する．

3 結果

いくつかの大学のサイトに対して本システムを実行した．その結果のひとつは以下のとおりである．

- サイトごとのデータの種類と大きさの比較
  図 1 に示すように，A のサイトは gif 画像が多いが，B のサイトは jpeg 画像が多いことが分かる．このように，サイトによって，提供しているデータの種類が様々で，そのデータが占める割合もサイトによって大きく異なることが分かる．

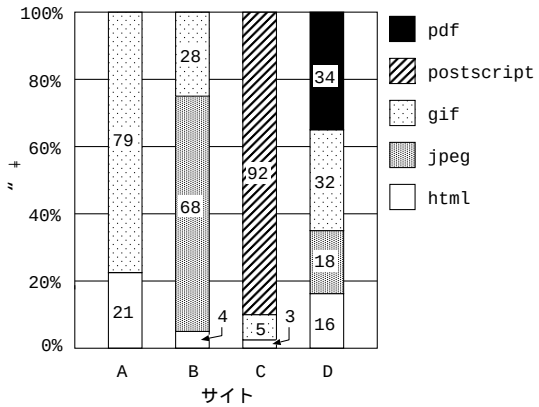

図 1: サイトごとのデータの種類の割合

4 おわりに

サイトごとに Web ページを収集し，それがどのようなデータの種類でどれくらいのデータ量であるか，様々な規格やブラウザでどれだけ対応している HTML ファイルがあるかの情報を提供するシステムを開発することができた．