DSIUシステムにおけるWebテキスト収集エージェント

2J－7

島津光伸　藤本和則

日本電信電話(株)コミュニケーション科学基礎研究所

1. はじめに

我々は、インターネット利用者に様々なアドバイスを与える「DSIU システム」の研究を進めている[1,2]。ここに、DSIU は"デシュウ"と読み、Decision Support for Internet Users の略である。DSIU はアドバイスのための推論知識を WWW 上のテキストから自動獲得する。このため、DSIU は、まず、知識獲得の源となるテキスト、特に「対象とする分野についての解説文が含まれるテキスト(以下、解説テキスト)」を WWW から収集する。DSIU システムにおいては、こうした解説テキストを高速に収集する枠組みの実現が重要となる。

WWW からテキストを収集するにあたっては、しばしば検索エンジンが利用される[3]。しかしながら、検索エンジンを直接利用する方法では、解説テキスト以外の URL も多く検索されてしまう。そこで、我々は、分野の重要な URL を集めた Hub ページを利用する研究を進めている。こうした Hub ページには、重要な URL のみが(人手で)集められている。したがって、こうした Hub ページに掲載される URL を集めることができれば、それをたどることにより多くの解説テキストを収集できるだろう。こうした Hub ページは、

・特徴的な語("リンク集"など)が用いられている。

・外部リンクが多く掲載されている。

という特徴をもつ。したがって、対象とするページが Hub ページであるかどうかを自動的に判定するのが比較的容易である。本稿では、DSIU の Web テキスト収集法として、Hub ページを自動的に見つけ出し、この Hub ページの URL をたどることによりテキストを収集する方法を提案する。

2. Hub ページ指向のテキスト収集法

提案の収集法の収集フローを下図に示す。

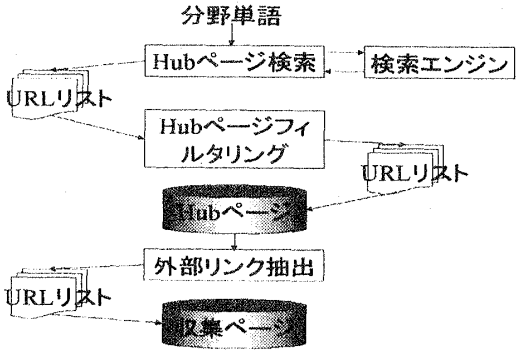

図 1.収集フロー

(1)Hub ページ検索

対象とする分野を示すキーワード(以下、分野単語)と、Hub 識別単語("リンク集"、"リスト"など)により、検索組を作成する。そして、この検索組をもとに検索エンジンを用いて、URL リストを取得する。

(2)Hub ページフィルタリング

HTML におけるハイパーリンクは大きく分けて、

・href="URL"の形式

(例 <a href="http://www.yahoo.co.jp/">)

・href="ファイル名"の形式

(例 <a href="/DSIU/www/top.html">)

の二つの形式がある。前者の形式のみを外部リンクとして判定し、抽出する。抽出された外部リンク数をカウントし、あらかじめ定められたリンク数よりも大きいページを Hub ページであると判定する。

(3)外部リンク抽出

Hub ページであると判定されたページの外部リンクを抽出し、このリンクの指すページを収集する。

"An Internet Search Robot Gathering Web-pages for Decision-making in DSIU systems", Mitsunobu SHIMAZU and Kazunori FUJIMOTO, NTT Communication Science Laboratories,2-4 Hikari-dai Seika-cho Souraku-gun Kyoto 619-0237 Japan.

3. 実験

3.1. 実験方法

対象とする分野「デジタルカメラ」について、本手法の収集速度に関する実験を行った。実験にあたっては、Hub 識別語として「ランキング」という単語を用いた。また、検索エンジンとしては goo[4]を利用した。Hub ページの判定にあたっては、20 以上の外部リンクを掲載するページを Hub ページとした。

[解説テキスト] デジタルカメラを解説するテキストをメーカの提供するホームページから集め、これを正解テキストとした。この正解テキストとして、92テキストを集めた。

[比較手法] 比較のため、デジタルカメラというキーワードをもとに、検索エンジンを直接使って収集する2つの手法を用意した。

- カテゴリ型検索エンジン(yahoo)[5] を用いた収集
  「デジタルカメラ」のカテゴリ以下に掲載される URL をたどりながら収集する手法
- ロボット型検索エンジン(goo)を用いた収集
  「デジタルカメラ」をキーワードとして検索された URL(上位1000)をたどりながら収集する手法

以上の条件のもとに、テキスト収集を行い、それぞれの手法でどのように正解の解説テキストが収集されるかを調べた。

3.2. 実験結果

前節の方法で収集を行ったときの、解説テキストの収集特性を図2に示す。図において、横軸は収集した全 URL 数、縦軸は収集できた解説テキスト数をそれぞれ表す。収集にあたっては、収集 URL 数が2000 になるまで行った。また、重複した URL は、カウントから除いた。

図2からわかるように、ロボット型検索エンジン(goo)を直接使った方法では、2000URLを収集した時点で、32の解説テキストしか集められていない。これに対し、提案の収集法では、2000URLを収集した時点で、48の解説テキストを集めることに成功している。(この 48 の解説テキストは、全体(92)の約51.1%である。)この

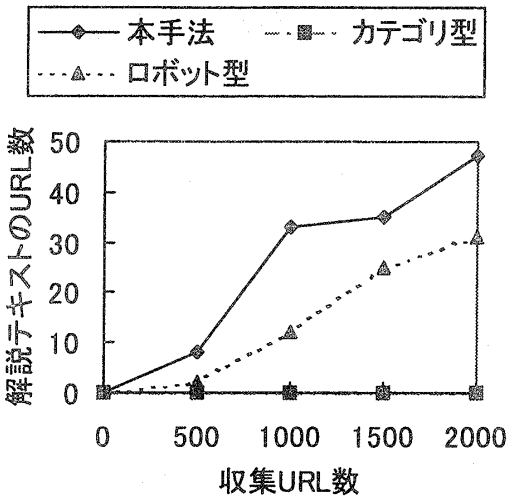

図 2. 解説テキストの収集特性

ように、Hub ページを利用して収集することにより、(Hub 識別語をうまく設定すれば)効率よく解説テキストを収集できることがわかる。なお、図からわかるように、カテゴリ型検索エンジン(yahoo)では解説テキストが収集されなかった。これは、今回正解の解説テキストとして用意したメーカのページは、デジタルカメラのカテゴリから数段下に位置づけられていたためと考えられる。

4. おわりに

本稿では、DSIU システムにおける Web テキスト収集法を提案した。そして、デジタルカメラに関する解説テキストを例に本手法の有効性を示した。なお、本実験では、検索エンジンとして goo を利用したが、他の検索エンジンを用いても有効であることを確認している。今後は、様々な分野へ適用したときの有効性を調べる予定である。

参考文献

[1] Kazunori Fujimoto and Kazumitsu Matsuzawa. Intelligent systems using webpages as knowledge base for statistical decision making. to appear in *New Generation Computing*,Vol.17,No.4,1999.
[2] 藤本和則,松澤和光. インターネット上の記述文から確率知識を構成する一手法: 構成の基本原理を中心に.情報学シンポジウム, pp. 129-136,1998.
[3] 村田剛志.参照の共起性に基づく発見手法.人工知能学会研究会資料 SIG-FAI-9901,pp.31-36,1999.
[4] Goo (http://www.goo.ne.jp)
[5] Yahoo (http://www.yahoo.co.jp)