2C-3

拡大アンカーテキストを利用し偏りにも考慮したフォーカスドクローラについて

遠藤雅樹† 高谷大二郎‡ 大野成義†

職業能力開発総合大学校 能力開発院情報通信ユニット†

職業能力開発総合大学校 電子情報システム工学科‡

1 はじめに

Google や Yahoo! などの検索エンジンは，クローラと呼ばれるプログラムを使って世界中の Web ページのデータを収集し，データベース化している．それに対してフォーカスドクローラ[1]は，特定のトピックに関するページのみを収集・データベース化するプログラムである．特定分野に関する検索エンジンを作りたいときや，キーワードで指定することが難しいページを収集したいときは上記の汎用の検索エンジンでは効率が悪い．本研究では人間のように拡大アンカーテキスト（アンカーテキスト及び周辺文字列）を判断材料としてリンクを取捨選択するフォーカスドクローラの開発を行う．さらに，収集したページの偏りを平準化するための機能を取り入れ，指定されたトピックに関するページをバランス良く収集できるようにする．

2 拡大アンカーテキスト

フォーカスドクローラの目的は，特定のトピックに関するページを選択的に収集することである．つまりフォーカスドクローラはクロール境界線を分析しなければならない．人間は，アンカーをクリックするとき，そのアンカーだけでなくアンカー周辺の文章も考慮して，リンク先のページに求めている情報があるかどうか判断している．羽田ら[2]は，この拡大アンカーテキストを利用してクロール境界線を分析するフォーカスドクローラを開発した．このクローラは，拡大アンカーテキストを形態素解析し，リンク先ページがトピックと類似しているか（ターゲットであるかどうか）をスコアとして計算し判定する．さらに，リンク先ページに関係する単語は，アンカーテキスト及びその直前・直後に最も多く存在することを調査により明らかにした．これを利用して，「重み」を単語に与え，リンク先ページのスコアを効果的に判定できるようにした．重みは，アンカーテキスト及びその前後は大きく，遠ざかるにつれて小さくしていく．これにより，リンク先ページに関係する単語を重要視することができる．

3 ターゲットか否かの判定方法

収集したページがターゲットであるか否か判定する方法に関して，ターゲットワードを与えておき，そのターゲットワードが出現するか否かで，そのページがターゲット否かを判断する方法がある[2][3]．一方，ページの特徴語を利用して判断する方法を採用しているクローラもある[4]．そこで予備実験として以下の判定方法の比較を行った．

方法１：ページに出現した単語を利用した判定．ページ中にターゲットワードのうち過半数が出現したらターゲットとする．

方法２：ページの特徴語を利用した判定．TF-IDF による特徴語の上位 25 ワードにターゲットワードのうちの過半数が出現したらターゲットとする．

上記の２つの方法のどちらの方の精度が良いか，人間による判定を正解として適合率と再現率で評価する．ターゲットワードの個数は 3,5,7 と変えてみた．また，人間による判定が必要なため，少ないが判定用収集ページ数は 50 とした．その結果が表１である．

方法２に比べて方法１の方が再現率は高いが適合率は低くなった．方法１は甘口判定で，方法２は辛口判定といえる．本来ターゲットであるページを誤ってターゲットでないと判定しないという意味では，方法１の方がフォーカスドクローラの判定方法として望ましいと考えられる．ただし，予備実験では特徴語を出現頻度で

表１ 判定方法に関する予備実験

		ターゲットワード数		
		3	5	7
方法１	適合率	0.60	0.50	0.86
	再現率	0.95	0.94	0.87
方法２	適合率	0.78	0.73	1.00
	再現率	0.82	0.42	0.33

An effectively focused crawling system using extended anchor text

†Masaki Endou, Shigeyosi Ohno. Unit of Information and Communication, Polytechic University

‡Daijirou Takaya. Department of Electronics and Information System Engineering, Polytechic University

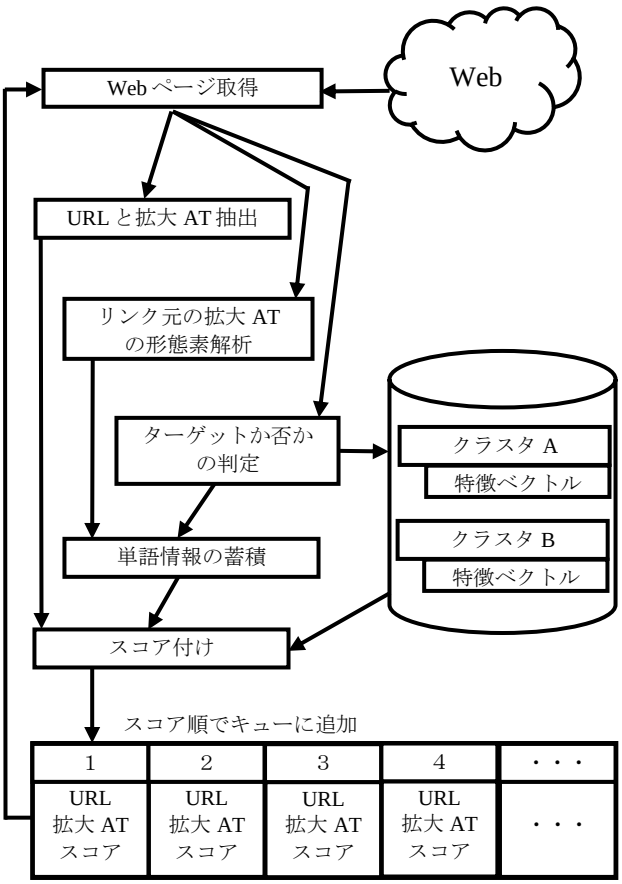

図１　処理の流れ

上位 25 ワードに限定し，判定に使用したページ数も 50 と少ない．より詳細な実験が必要である．

3　ページのクラスタリング

フォーカスドクローラの問題点として，収集した Web ページに偏りが生じるというものがある[4]．多くの場合，既に多数集められた Web ページに類似する Web ページを更に収集するより，既に収集された Web ページと異なる Web ページを収集することが望まれる．そこで，ある程度 Web ページを収集したら，ページをいくつかのクラスタに分ける．このとき１つのクラスタにページが多く集まっている場合，同じような内容のページばかりを収集していると考えることができる．このことからトピックに関連するページの偏りを解消することを，全てのクラスタが平等に収集されることと定義する．偏りを平滑化するために，求めた各クラスタを用いて新たに特徴ベクトルを生成する．十分に収集できていると判別されたクラスタがあった場合，そのクラスタの特徴ベクトルと類似している Web ページのクロール優先順位を下げる．本研究では重みを操作して，スコアの値を調整する．これを一定数クロールするたびに繰り返して，偏りを平滑化する．拡大アンカーテキストの利用とクラスタリング機能を踏まえて，提案手法の処理の流れを図１に示す．図中の拡大 AT とは拡大アンカーテキストのことである．

4　実験とスタートセットについて

羽田らのクローラは Web ページの偏りを考慮していなかった．そこでクラスタの機能を付加し，ターゲットか否かの判定方法も精度を上げたクローラを提案する．そのプロトタイプを試作し，小規模ながら実験を行ったので報告する．

フォーカスドクローラは一般的なクローラと同様に取得したページから張られているリンクを辿ることでページを収集する．違いは辿るリンク先のページ内容を（拡大）アンカーテキストを利用して推測し，リンクを辿る優先順位を変えることにある．推測の精度がフォーカスドクローラの性能の良さを左右することになるが，そもそもリンク先に収集したいページが存在しなければ，推測の精度がいくら良くても意味がない．そこで，シードページの集合（スタートセットと呼ぶ）をどのように選ぶかは重要になってくる[5]．

実験では，幅広い分野のページをカバーするために，ディレクトリ検索である yahoo と goo のトップページ，辞書サイトの Wikipedia の「一覧の一覧」とニコニコ大百科の「一覧の一覧」のページ，ブログサイトの Goo ブログの「ジャンル一覧」とアメーバブログの「ジャンル一覧」のページ，そしてニフティの検索エンジンをスタートセットとして採用した．

5　参考文献

[1]S. Chakrabarti, M. van den Berg, B. Dom, “Focused crawling:a new approach to topic-specific Web resource discovery”, Computer Networks, 31(11-16), pp.1623-1640 1999.
[2]羽田哲也，大野成義，寺町康昌，石川博，”重み付き拡大アンカーテキストを用いたフォーカスドクローラの開発”，情報処理学会 DBS 研究会報告，65，pp.491-496，2007.
[3]富山北斗，伊藤栄典，廣川左千雄，”自己学習型トピッククローラーの構築と評価”，DEWS 2006，3B-i11.
[4]上村祐輝，北須賀輝明，有次正義，”トピックに関連する Web ページの偏りを考慮した Focused Crawler”，DEIM Forum 2011 F8-5.
[5]大村啓明，陳漢雄，古瀬一隆，”サーチエンジンを用いた Focused Crawling に関する研究”，情報処理学会第 72 回全国大会，3R-4，2010.