社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143（83）
2007／7／4

# 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ーの開発

羽田 哲也† 大野 成義†† 寺町 康昌†† 石川 博†††

† 職業能力開発総合大学校 電気・情報専攻 〒229–1196 神奈川県相模原市橋本台 4–1–1
†† 職業能力開発総合大学校 情報システム工学科 〒229–1196 神奈川県相模原市橋本台 4–1–1
††† 静岡大学 情報学部情報科学科 〒432-8011 浜松市城北 3-5-1
E-mail: †m18520@uitec.ac.jp, ††{ohno,teramati}@cs.uitec.ac.jp, †††ishikawa@inf.shizuoka.ac.jp

**あらまし** キーワードで指定できないような特定分野の情報を大量・網羅的に収集するような要求に汎用的な検索エンジンは応えられない。一方、人間は Web 上から必要な情報を探し出す際、アンカーテキストやその周辺の文字列などからリンク先のページに求めている情報があるかどうかを判断している。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その周辺文字列）を判断材料としてリンクを取捨選択し、特定分野の Web ページのみを大量・高速に収集するクローラーの開発を目指す。また、精度の向上を図るために、拡大アンカーテキストに重みを付けることを検討する。

**キーワード** Web とインターネット，情報検索，フォーカスドクローラー，アンカーテキスト

# Focused Crawling Using Weighted Extended Anchor Texts

Tetsuya HADA†, Shigeyosi OHNO††, Yasuaki TERAMACHI††, and Hiroshi ISHIKAWA†††

† Department of Information System Engineering, Polytechnic University Hashimotodai 4–1–1, Sagamihara-shi, Kanagawa, 229–1196 Japan
†† Department of Information System Engineering, Polytechnic University Hashimotodai 4–1–1, Sagamihara-shi, Kanagawa, 229–1196 Japan
††† Faculty of Information, Shizuoka University Jouhoku 3–5–1, Hamamatsu-shi, 432–8011 Japan
E-mail: †m18520@uitec.ac.jp, ††{ohno,teramati}@cs.uitec.ac.jp, †††ishikawa@inf.shizuoka.ac.jp

**Abstract** A general-purpose search engine cannot satisfy the demand which collects the information on topical fields in large quantities and comprehensively. The topical fields are specified not using keywords, but using exemplary documents. On the other hand, when the user obtains the required information from Web, he judges whether there is any information for which page of the link place is asked from the anchor text and strings befor and after it. In this paper, we develop focused crawler that more quickly selectively seeks out a lot of pages of topical fields by making an extended anchor text (that include nouns before and after the link URLs) so that people selects the link and seeks out pages. Moreover, in order to improve precision, we discuss about weight on extended anchor text.

**Key words** Web and Internet, Information retrieval, Focused Crawler, Anchor Text

## 1. はじめに

インターネット上の情報は年々増加しているが、特に近年は Blog など動的に生成される Web ページの急増に伴い、情報量は爆発的に増加している。静的な Web ページだけで 150 億ページ、動的な Web ページまで含めると 350 億ページを超えるとも言われているが、個人が必要とする情報量はそれほど多くはない。情報爆発と呼ばれる状況において、個人が本当に必要としている情報を探し出すために、Google [1] や Yahoo [2] といった汎用的な検索エンジンがよく利用されている。しかし、このような汎用目的の検索エンジンでは全ての人が満足する検索結果を返すことは不可能である。また、詳細な検索要求（例えば、あまり一般的ではない専門用語を用いた検索など）に対応できるとは限らない。このため、論文を検索対象とする Google Scholar [3] をはじめとして、特定分野の情報だけに特化した検索サービスを提供している専門検索エンジンも多数存在している。更に、必要な情報が載っている Web ページを探し出すだけでなく、インターネット上に存在する情報から自動的に知識を発見する Web マイニングの研究も盛んに行われている。こういった専門検索エンジンの構築や Web マイニングの研究の際には、対象となる特定分野の高品質なデータを多数、収集しなければならない。そのため、特定分野の Web ページを効率良

く高速に収集するシステムが必要である。また、シラバスのような、キーワードにより指定することが難しい Web ページを収集するためにも必要である。

一方、情報探索の際には検索エンジンだけでなく、Yahoo や Excite [4] に代表される Web ディレクトリもよく利用される。Web ディレクトリにおける Web ページの分類は人手で行われており、自動分類しようとする研究が盛んに行われてきた。従来は、ターゲットページから抽出した単語を用いて自動分類してきたが、近年注目されているのは、ターゲットページにリンクしているページ（リンク元ページ）に含まれる単語を用いて分類する方法である。この方法は、Web ページの収集の際に特定分野の Web ページのみを収集する為に利用できる。

人間はアンカーテキストやその周辺の文字列などから、リンク先のページに求めている情報があるかどうかを判断している。簡単な例を挙げると、Yahoo のトップページには最新ニュースの見出しが掲載されており、見出しにはニュース本文が掲載されているページへのリンクが設定されている。もし、見出しを読み、興味がある内容でありそうならばリンクをクリックするだろうし、興味がない内容であるようならリンクをクリックしたりはしないだろう。Google などでキーワードを入力し検索した際も、検索結果として羅列される Web ページの文章を読み、必要な情報がリンク先にあるかどうか判断している。

本研究では、人間の情報探索と同じように、アンカーテキスト及び周辺文字列（拡大アンカーテキスト）を判断基準にリンク先のページに必要な情報があるかどうかを推測し、また、拡大アンカーテキストに重みを付ける事により、極力不要な Web ページを収集することなく、特定分野の Web ページのみを収集するクローラーの開発を目的とする。

以降、2 章でクローラー及びアンカーテキストの関連研究について、3 章では予備実験について、4 章で拡大アンカーテキストについて、5 章で提案手法、6 章で今後の予定について述べる。

## 2. 関連研究

### 2.1 クローラー関連研究

まず特定分野の Web ページのみを収集するクローラーに関する研究事例をあげる。Chakrabarti [5], [6] らは、収集したい特定分野に関連するページから収集を始め、関連する Web ページを選択的に収集していくフォーカスドクローラー について研究している。例えば『安倍晋三』に関する Web ページを集めたいという場合、政治に関するページの近くに多く存在していると経験的に分かるはずである。つまり、関連度の高いページ同士はリンクで繋がっている可能性が高く、関連度の高いページを集めていけばターゲットも集まりやすいだろう、というのが、このクローラーの発想である。ちなみに、関連度は、事前によく学習された文書分類器で計算する。

Diligenti [7] らは、ターゲットページ周辺の典型的な文脈（リンク情報）を学習した context graphs を用いたクローラーについて研究している。図 1 に context graph の概念図を示す。

ターゲットページに 1 回のリンクで行けるページ群をレイヤー 1、2 回のリンクで行けるページ群をレイヤー 2 とし、同様に、レイヤー 3、レイヤー 4 と定義する。こうして、ある一定の大きさをもったグラフ（context graphs）を事前に作成する。このグラフはターゲット周辺の特徴を示した地図のようなものであり、地図を用いながら探索することで、ターゲットページを発見しやすくなるという発想である。

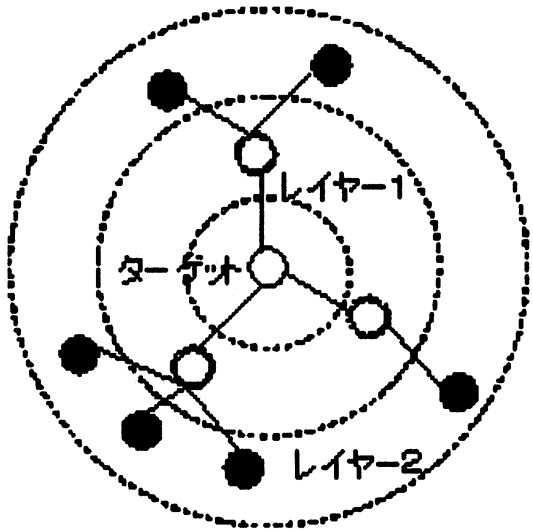


図 1　context graphs の概念図

Ester [8] らは、Web サイトを発見する外部クローラーと、サイト内をクローリングする内部クローラーを統合した Web クローラーの研究を行っている [11]。この手法は、サイト内のページを探すのではなく、トピックに関するサイトを探すことを目的としている。

冨山 [9] らは、リンク先ページ判定関数とリンク元ページ判定関数の 2 つの判定関数を持つ自己学習型トピッククローラー（G-CRAWLER）を試作している。G-CRAWLER の処理の流れを図 2 に示す。リンク先ページ判定関数はリンク先ページがターゲットかどうかを判定し、その結果を元にリンク元ページ判定関数はリンク元ページから抽出した単語情報を正例もしくは負例としてデータベースに登録する。そして、抽出したリンクにデータベースにある単語情報を元にスコアリングし、スコアの高い順から収集していく。

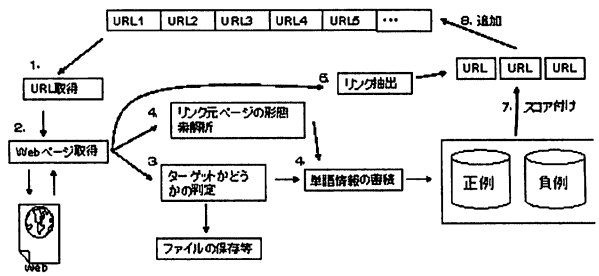


図 2　G-CRAWLER の処理の流れ

### 2.2 アンカーテキスト関連研究

次に、リンク元の Web ページ中の単語を用いてリンク先の Web ページの分類を行う研究事例をあげる。これらの研究は、Web ディレクトリ用に人手でなく自動分類しようという目的であり、分類対象のページは収集済みである。そのため、収集しながら分類するフォーカスドクローラーの開発とは異なるが、リンク元ページの単語を用いるという意味で共通性がある。

Bulms [10] らはリンク元ページのアンカーテキストのみを用いてリンク先ページを分類しているが，リンク元ページ全体のテキストを分類に使う場合と比べ、精度にあまり差は無かった。Chakrabarti [11] や Furnkranz [12] らは、アンカーテキストを含む段落全体やリンク元ページの見出しを使って分類しているが，段落が長くなるにつれ分類精度は悪くなっていた。Glover [13] らは、リンク元ページのアンカー前後２５単語を拡張アンカーテキストと定義し、アンカーテキスト、拡張アンカーテキスト、全テキストを比較検討し、拡張アンカーテキストが最も精度がよいという結果を示した。

大坪 [14] らは、Web ページの DOM 構造を解析したうえで、アンカーテキストおよびその周辺文字列を LSP(Local Semantic Portion)、ページタイトルや見出しを USP(Upper-level Semantic Portion) として抽出し分類に用いる手法を提案し、Glover らの拡張アンカーテキストよりも精度が良いことを実験で示した。しかし、大坪らの研究では、英語の Web ページが対象であり、日本語の Web ページにおいても USP や LSP が有効かどうか検証する必要がある。

神林 [15] らは、拡大アンカーテキストを用いて Blog のクラスタリングを行う手法を提案している。

## 3. 事前調査

### 3.1 調査方法

リンク先ページと関連する単語がリンク元ページのどこに存在することが多いのか、また、日本語の Web ページにおいて、大坪らの提案している LSP や USP がどの程度有効なのか、事前調査を行った。表 1 に事前調査で調べた Web ページ数を載せる。

| | 公的 | 私的 | 企業 |
|---|---|---|---|
| ターゲット | 10 | 3 | 4 |
| リンク元 | 50 | 15 | 20 |
| 合計 | 85 | | |

表 1　事前調査の Web ページ数

まずターゲットページとして、政府機関などの公的な Web ページや個人が運営する私的な Web ページ、企業の Web ページを計１０ページ収集した。1 つのターゲットページに対しリンク元ページを 5 ページに限定し、これらのターゲットページそれぞれに対するリンク元ページを Google のリンクページ検索を用いて集め、アンカーテキストおよびその周辺文字列やページタイトル、見出しについて調査した。

### 3.2 結　　果

調査結果を表 2, 4 に示す。

| | 関係有 | 全体 |
|---|---|---|
| LSP | 80 | 85 |
| USP | 5 | 27 |

表 2　LSP と USP の個数

| アンカーテキスト | 62 |
|---|---|
| 直前・直後 | 16 |
| その他 | 2 |

表 3　関係単語の位置

ほぼ全てのリンク元ページに LSP が存在しており、リンク先ページの内容に関係しているのに対し、USP が存在していたのはほぼ半数であり、リンク先ページに関係しているのはわずかに 5 ページのみであった。調査数が少ないが、日本語の Web ページにおいては USP はさほど重要ではない可能性が高い。

また、LSP において、リンク先ページに関連する単語は同じ文章内に存在し、特にアンカーテキスト自体に多く含まれており、それ以外では、アンカーテキストの直前・直後に多く見られた。別の文章など、アンカーテキストから離れているところにはほとんど見受けられなかった。

次に、ターゲットページへのリンクが存在した形式を表 4 に示す。

| | |
|---|---|
| リスト | 10 |
| テーブル | 18 |
| 段落 | 22 |

表 4　調査結果：形式分類

リスト (< ol >, < ul >, < dl >) やテーブル (< table >) は多数の Web ページへのリンクを集めたリンク集において、よく見かける形式であり、リンク先ページの名前や URL など、必要最低限の情報しか存在しなかった。また、アンカーテキスト内にリンク先ページに関係する単語が存在したのは、ほとんどがテーブルやリスト形式のリンク集ページであった。逆に段落形式 (< p >) の場合、日記などの話の流れの中でリンク先ページを紹介しているため、文章で記述されている。そのため、段落形式ではアンカーテキストの直前や直後にリンク先ページに関係する単語が存在していた。

### 3.3 考　　察

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し、離れたところにはほとんど見受けられなかったことから、アンカーテキストから遠ざかっていくごとにリンク先ページへの関連性は弱くなっていくことが分かる。つまり、アンカーテキストおよびその前後は重みを大きく、遠ざかるにつれ重みを小さくしていくことにより、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、特定分野の Web ページだけを収集していくことが可能になると考えられる。そこで、重みの付け方として、以下の (1) 式, (2) 式, (3) 式いずれかの利用が考えられる。これらは、太田らのクラスタリングに関する研究 [16] で定義・使用されており、NTCIR-4 におけるクラスタリングのコンテストで高成績を収めている方法である。

(1) 式, (2) 式, (3) 式はそれぞれ単語 $f$ の重みを定義しており、(1) 式は線形的に減少する値、(2) 式, (3) 式は曲線的に減少する値となる。$T$ は総形態素数、$p$ は文書内出現位置である。(1) 式の $a,b$ は任意の定数であり、ここでは $a=50,b=0.5$ に設定している。総形態素数が 10 のときの (1) 式, (2) 式, (3) 式のグラフを図 3 に示す。

$$lw(p) = \begin{cases} 1 - \frac{bp}{a} & a > bp \\ 0 & otherwise \end{cases} \quad (1)$$

$$sw(p) = \sin\left(\frac{T-(p+1)-1}{2\times T}\pi\right) \quad (2)$$

$$esw(p) = \sin\left(\frac{T+(p+1)-1}{2\times T}\pi\right) \quad (3)$$

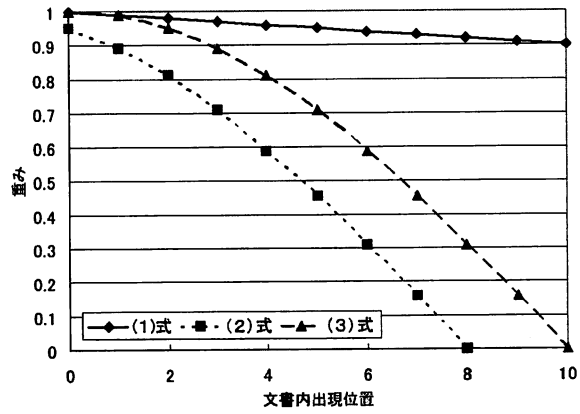


図 3　重みのグラフ

## 4. 拡大アンカーテキスト

ある Web ページから他の Web ページにリンクする際、アンカーテキストにリンク先ページに関係する単語や文章を書く人もいれば、アンカーテキストには URL しか書かず、リンク先ページの説明はアンカーテキストの前後の文章で行う人もおり、Web ページによってバラバラである。そのため、アンカーテキストだけでリンク先の内容を推測するのは難しく、アンカーテキスト周辺の文字列も参照する必要がある。

そこで、事前調査の結果から、HTML タグを除去したアンカーテキスト及びその周辺文字列を拡大アンカーテキストと定義する。

前節 3.3 で示した図 3 において、アンカーテキストを文書内出現位置 0 とした場合、アンカーテキストから前後両方向に離れるにつれ、重みが減少していくことになる。事前調査により、リンク先ページに関係する語はアンカーテキストとその直前・直後にあることが多く、離れているところに存在する語の重要度は低いといえる。徐々に重みが減少する (1) 式より、急激に重みが減少する (2) 式, (3) 式の方が拡大アンカーテキストへの重み付けに向いていると考えられるため、提案手法では (2) 式, (3) 式を採用する。

## 5. 提案手法

図 4 に提案手法の処理の流れを示す。図中の拡大 AT とは拡大アンカーテキストのことである。

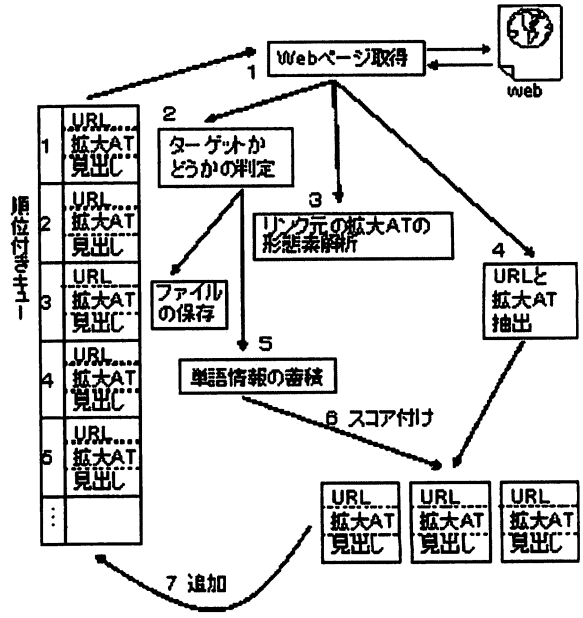


図 4　本手法の処理の流れ

優先順位付きキューから URL を取り出し、その Web ページに HTTP アクセスし、HTML ソースを取得する。その Web ページがターゲットページかどうか判定するとともに、リンク元の拡大アンカーテキストの形態素解析を行い、単語情報を蓄積する。単語情報の蓄積は、以下の 4 式、5 式を元に行う。

$$P_t(f_i) = P_{t-1}(f_i) + w(p) \quad (4)$$

$$N_t(f_i) = N_{t-1}(f_i) + w(p) \quad (5)$$

ここで、$f_i(i = 0, 1, ...)$ は拡大アンカーテキストに出現する単語、$P$ はターゲットに結びつく単語情報 (Positive Word) を格納した配列、$N$ はターゲットに結びつかない単語情報 (Negative Word) を格納した配列である。また、$w(p)$ は前述の拡大アンカーテキスト位置 $p$ にある単語 $f_i$ の重みを表し、$t$ はクローラーが辿るページの訪問順番である。

同時に、HTML ソースを解析し、他の Web ページへのリンクとその拡大アンカーテキストを取得する。そして、以下の式を用いて、蓄積した単語情報を元に、取得した URL へのスコア付けを行う。

$$score(URL) = \sum \frac{w(p)\times P_t(W_x)}{w(p)\times P_t(W_x) + w(p)\times N_t(W_x)} \quad (6)$$

ここで、$\sum$ は、拡大アンカーテキストの最初から最後まで全ての位置について和をとる。位置が異なれば、同じ単語につい

て複数回加算することになる。スコア付けされたURLはリンク元の拡大アンカーテキストともにキューへ追加され、次の探索のためにキューから最もスコアの高いURLを取り出し、同様の処理を繰り返していく。

提案手法の処理手順は2.1クローラー関連研究で紹介した富山らのG-CRAWLERとほぼ同じであるが、提案手法ではリンク元の拡大アンカーテキストを元にスコア付けしており、G-CRAWLERではリンク元ページ全体を元にスコア付けしている。その違いを図5に示す。

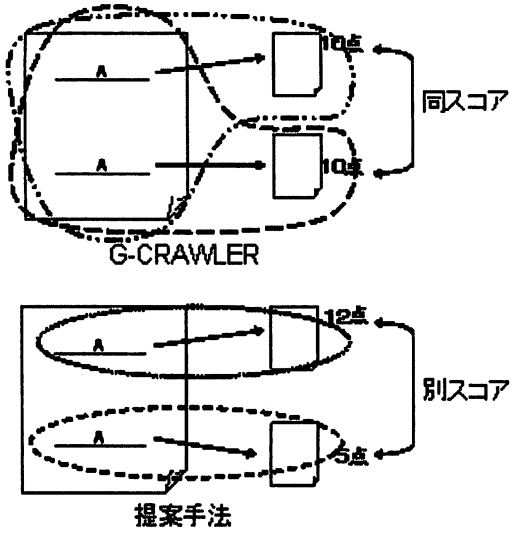


図5　スコア付けの違い

G-CRAWLERではリンク元ページ全体の形態素解析を行い、その単語情報を蓄積し、URLへのスコア付けに用いている。そのため、1つのページに多数のリンクがあった場合、それらのスコアは同じ値になってしまう。つまり、1つのリンク元ページに対し複数のリンク先ページ、という1対多の関係である。G-CRAWLERがターゲットにしているのは、シラバスやレシピ集のような、1つのページに多数のページへのリンクがあるリンク集である。イメージとしては図6のように、網で一気にWebページを獲るのだが、必要なWebページ（=魚）以外にも不要なWebページ（=木の枝、貝）まで獲ってしまう。

一方、提案手法ではリンク元の拡大アンカーテキストの形態素解析を行い、その単語情報を蓄積し、URLへのスコア付けに用いている。つまり、それぞれの拡大アンカーテキストからスコアを算出するため、1つのページに多数のリンクがあったとしても、それらのスコアは全て同じにはならない。リンク元の拡大アンカーテキスト1つに対し1つのリンク先ページ、という1対1の関係になる。イメージとしては図7のように、不要なWebページ（=木の枝、貝）には見向きもせず、必要なWebページ（=魚）だけを狙って釣り上げていく。

ちなみに、正しく書かれていないHTMLソースにも対応出来るよう、HTMLソースの解析には、HTMLタグの不備を補完する機能を持っている“CyberNeko HTML Parser”を使用する。

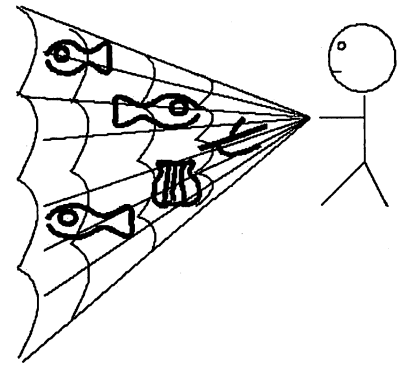

図6　G-CRAWLERのイメージ

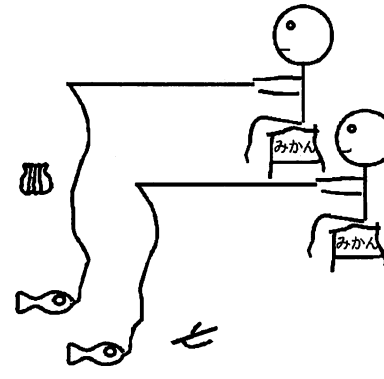


図7　提案手法のイメージ

## 6. おわりに

本論文では、拡大アンカーテキストを用いて特定分野のWebページのみを収集するクローリング手法を提案した。また、精度の向上を図るために、拡大アンカーテキストに重みを付けることを検討した。今後は、提案手法を採用したクローラーを使った実験を通して、性能評価、他の手法との比較を行う。また、ページタイトルや見出しの寄与は小さいが、無視は出来ない。ページタイトルや見出しの寄与をより詳細に定量的に評価し、精度向上を図る。

### 文　献


[1] Google : http://www.google.co.jp/
[2] Yahoo! : http://www.yahoo.co.jp/
[3] Google Scholar : http://scholar.google.com/
[4] Excite : http://www.excite.co.jp/
[5] Soumen Chakrabarti et al., “Focused crawling: a new approach to topic specific Web resource discovery”, Computer Networks, Vol31, No.11-16, pp.1623-1640, 1999.
[6] Soumen Chakrabarti et al., “Accelerated focused crawling through online relevance feedback”, Proceedings of the 11th inrnational conference on World Wide Web, pp.148-159, 2002.
[7] M. Diligenti et al., “Focused Crawling Using Context Graphs”, Proc. of the 26th International Conference on Very Large Data Bases(VLDB2000), pp.527-534,2000.
[8] Martin Ester, Hans-Peter Kriegel, Matthias Schubert, “Accurate and Efficient Crawling for Relevant Websites”, Proc. of the 30th VLDB

Conference (VLDB2004), pp.396-407, 2004.
[9] 冨山北斗, 伊東栄典, 廣川佐千男：“自己学習型トピッククローラーの開発と評価”, DEWS2006, March.2, 2006.
[10] Avrim Blum, Tom Mitchell：“Combining Labeled and Unlabeled Data with Co-Training”, COLT98, pp.92-100, 1998.
[11] Soumen CHakrabarti, Byron Dom, Piotr Indyk：“Enhanced hypertext categorization using hyperlinks”, SIGMOD'98, pp.307-318, 1998.
[12] Johannes Furnkranz：“Exploiting Structural Information for Text Classification on the WWW”, IDA'99, pp.487-497, 1999.
[13] Eric J.Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M.Pennock, Gary W.Flake：“Using Web Structure for Classifying and Describing Web Pages”, WWW2002, pp.562-569, 2002
[14] 大坪正範, Bui Quang Hung, 土方嘉徳, 西田正吾：“アンカー関連テキストを用いたWebページ分類方式の設計と実装”, WI2-2006, pp.1-6, 2006.11.
[15] 神林 真実, 福田 直樹, 石川 博：“Blog空間における拡大アンカーテキストと明示的リンク解析に基づくクラスタリング手法”, DEWS2007, March.2, 2007.
[16] N.Ohta, H.Narita, S.Ohno：“Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 WEB Task”, Proceedings of NTCIR-4, 2004.
[17] 鎌田 基之, 福田 直樹, 石川 博：“Trckbackと特徴語に基づくBlogクローリングとBlog記事の推薦”, DEWS2007, March.2, 2007.