# サーチエンジンを用いたFocused Crawlingに関する研究*

大村 啓明†　陳 漢雄‡　古瀬 一隆§

†,‡,§ 筑波大学 システム情報工学研究科　〒 305-8577 茨城県つくば市天王台 1-1-1

## 1 はじめに

現在 web 空間上には膨大な数の web ページが存在する。このような膨大なページからは、ある特定の分野の web ページを大量に取得し、データベース化することで、新たな知識発見が行えると考えられる。特定の分野の web ページを取得するには Focused Crawler と呼ばれる Crawler が使われている。通常の Crawler と Focused Crawler の異なる点は、通常の Crawler は取得したページの内容を考慮しないが、Focused Crawler は考慮する事である。Focused Crawler は取得したページに対してユーザが指定した分野との類似度を求め、類似度が高い場合にはさらに crawl を行い、類似度が低いページは crawl を行わない。これによって特定の分野のみのページを crawl する事が出来る。既存の Focused Crawler の問題として、指定した分野のページが近くに存在しない場合、多くの無駄なクロールを行わなければいけないことがあげられる。

本稿では、この問題に対して取得ページの正解率が減少したときに、サーチエンジンの検索結果を用いてシードページを変更し、Focused Crawler が効率良くページを取得する方法を提案する。

## 2 Focused Crawler

Crawler とは web ページ中のリンクを順次辿ることにより、web ページを網羅的に取得し、データベース化するプログラムである。主にサーチエンジンのデータベースや、統計調査に使われる。

### 2.1 Focused Crawler の動作

Focused Crawler の動作を説明する。図 1 で、矢印は web ページからのリンク、□のアルファベットは web ページ名、□の番号は取得する順番を表す。また、□の左上の数字は優先度を表す。

始めに seed page から URL を抽出し、各 URL の優先度を計算し、図 2 に示した priority queue に記録する。Focused Crawler は priority queue から優先度の最も高い URL の web ページを取得する。これらの動作を繰り返し行うことで特定の分野の web ページを取得していく。優先度の計算はよく学習された文書分類器などが使用される。

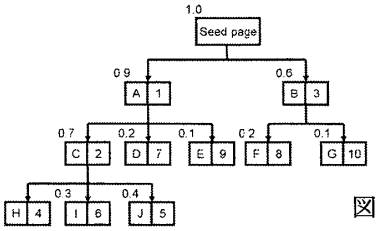


図 1: Focused Crawler

| | | |
|---|---|---|
| 1 | A | 0.9 |
| 2 | C | 0.7 |
| 3 | B | 0.6 |
| 4 | H | 0.5 |
| 5 | J | 0.4 |

図 2: priority queue

*Focused Crawling Mechanism Exploiting Search Engine Results
†Hiroaki Ohmura, Graduate School of SIE, University of Tsukuba
‡Hanxiong Chen, Graduate School of SIE, University of Tsukuba
§Kazutaka Furuse, Graduate School of SIE, University of Tsukuba

### 2.2 Focused Crawler の問題点

Focused Crawler は一般的な crawler 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 crawl していく。このため、取得したページから有効なページにリンクが張られていない場合、数多くの無駄なページを取得しなければいけない問題がある。

優先度を求めるための手法は様々な方法が提案されているが、この問題に対してこれまでの手法では対応していない [1][2]。

この問題の原因はシードページが一度しか与えられていないためである。この問題に対して、本稿ではサーチエンジンを用いて新たなシードページを与える方法を提案する。

## 3 提案手法

まず、提案手法の考え方を説明する。提案手法は事前学習フェーズ、Crawl フェーズの二つのフェーズから成り立つ。事前学習フェーズでは各リンクの優先度を決定するときに使用する優先度配列を求める。crawl フェーズでは、優先度配列を元に crawl を行う。

### 3.1 事前学習フェーズ

始めに crawl フェーズで crawl したページに含まれる URL の優先度を決定するために優先度配列を作成する。Focused Crawler はこの優先度配列に含まれる単語がアンカーテキストに存在した場合、有効な URL だと判断される。優先度配列の作成手順を図 3 に示す。

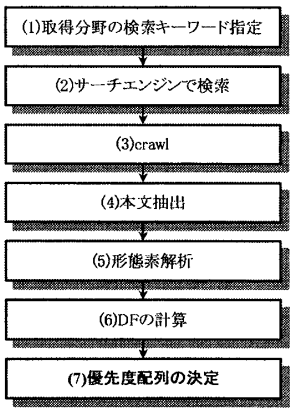


図 3: 事前学習フェーズ

手順

(1) 始めに取得したい分野の検索キーワードの指定を行う。
(2) サーチエンジンを使って指定した検索キーワードで検索を行う。
(3) 検索結果の上位 n 件の URL を crawl する。
(4) 取得した web ページのタグ情報などを除去し、本文を抽出する。
(5) 抽出した本文を形態素解析にかけ、名詞の単語だけ抽出する。

(6) 抽出したそれぞれの単語の DF を計算する。

(7) DF のスコアが高い上位 m 件を優先度配列とする。

## 3.2 Crawl フェーズ

Crawl フェーズでは優先度配列に基づいて優先度を決定し、crawl を行う。図 4 に手順を示す。

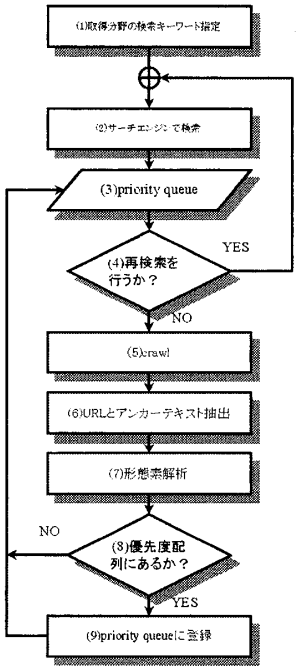


図 4: crawl フェーズ

手順

(1) 始めに取得分野の検索キーワードを指定する。この検索キーワードは事前学習フェーズのキーワードと同じである。

(2) サーチエンジンを使って検索キーワードの検索を行う。一周目は与えられた検索キーワードで検索を行う。二周目以降の検索キーワードは、初めに指定した検索キーワードと優先度配列の中で検索キーワードになっていない最も優先度の高い単語の and 検索とする。

(3) サーチエンジンの結果を priority queue に代入する。優先度は 0-1 の範囲で任意の実数を取る。一周目に与えられたシードページの優先度は 1 とする。

(4) priority queue を見て、再検索を行うか判断する。判断基準は、取得したページの正解率が一定以下であった場合と、まだ取得していないページが存在しなかった場合である。

(5) priority queue を見て、最も優先度の高い URL を取得する。

(6) 取得したページの URL とアンカーテキストを抽出する。

(7) それぞれのアンカーテキストを形態素解析にかけ、名詞の単語だけ抽出する。

(8) 抽出した名詞の単語が優先度配列に存在する場合には、その単語の優先度が URL の優先度となる。これを全てのアンカーテキストに対して行う。また、複数の単語が一つのアンカーテキストに存在した場合は、最も高い値が優先度となる。優先度配列に存在しない場合は再び crawl を行う。

(9) URL と優先度を priority queue に登録し、再び crawl を行う。

# 4 実験

## 4.1 実験概要

今回の実験では、取得する分野を“baseball”,“music”,“yankees”とし、幅優先 crawler との比較実験を行った。検索には Yahoo! API を使用した。事前学習フェーズでは検索結果の上位 100 ページを crawl し、優先度配列は上位 1000 単語とした。Crawl フェーズでは検索結果の上位 10 ページをシードページとした。

また、Beyesian filter は正解ページを 50 ページ、不正解ページを 500 ページを学習させた。ストップワードは指定した分野の検索キーワード以外の 10 個の検索キーワードで事前学習フェーズを行い、2 つ以上の検索キーワードで優先度が 0.3 以上となった単語をストップワードとした。本稿では、”baseball”の結果のみを載せる。

## 4.2 実験結果

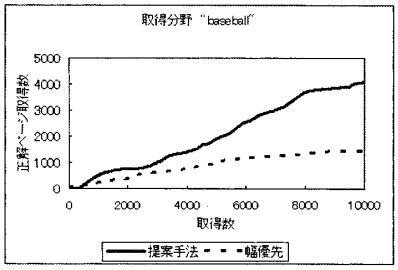


図 5: baseball

実験結果を図 5 に示す。図 5 の横軸は crawler が取得したページ数、縦軸は正解ページの取得数を表す。幅優先 crawler はシードページを変更し、3 回の平均を取ったものである。図 5 の提案手法の結果を見てみると、継続的に正解ページを取得できていることが分かる。この結果より従来の Focused Crawler の問題を解決出来ていることが分かる。

また、図 5 を見て分かるように提案手法が幅優先 crawler を大きく上回っていることから提案手法が幅優先 crawler よりも効果があることを示すことが出来た。

# 5 まとめと今後の課題

本稿では、従来の Focused Crawler の問題を解決する方法として、サーチエンジンを用いた Focused Crawling の提案を行った。今回の実験によって、既存の Focused Crawler の、取得したページの近くに指定した分野のページが存在しない問題を解決することが出来た。今後は取得するページを増やした場合の結果の考察、既存の Focused Crawler との比較、提案手法の Focused Crawler がどの程度の規模の分野に対して有効であるかの確認を行う。

# 参考文献

[1] S. Chakrabarti, M. van den Berg, B. Dom, ”Focused Crawling: a new Approach to Topic-Specific Web Resource Discovery”, In Proceedings of the 8th International WWW Conference, May 1999.

[2] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles, M. Gori, ”Focused Crawling Using Context Graphs”,In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000), September 2000.