自己学習型トピッククローラーの開発と評価

冨山 北斗† 伊東 栄典‡ 廣川 佐千男‡

†九州大学システム情報科学府 〒812-8581 福岡市東区箱崎 6-10-1
‡九州大学情報基盤センター 〒812-8581 福岡市東区箱崎 6-10-1
E-mail: †hokuto.tomiyama@i.kyushu-u.ac.jp, ‡{itou, hirokawa}@cc.kyushu-u.ac.jp

あらまし Web 上の情報量は、個人が必要とする情報量をはるかに上回っている。利用者の情報収集を支援する、Google や Yahoo 等の汎用的な検索エンジンが開発されている。しかし汎用的な検索エンジンでは、特定の分野について網羅的に収集し、情報をまとめるといった要求には応えられない。著者らは、特定分野の Web ページ収集を効率的に行なうトピッククローラーの研究開発を行なっている。トピッククローラーでは、トピックに関するページの判定精度も必要であるが、トピックページへ早く辿りつく速度も重要である。著者らは、One man & his dog システムと呼ぶトピック判定とリンク選出戦略機能を連携させるシステムを試作した。また、２つのトピックについての収集実験を行い本論文の手法の効率性を調査した。

キーワード Web とインターネット，情報検索，知識発見，トピッククローラー

1. はじめに

Web（WorldWideWeb）の情報量は個人が必要とする情報量をはるかに上回っており、情報爆発と呼ばれる状況を作り出している。そこで、利用者が求める情報を Web から探し出すために、Google[1]や Yahoo[2]といった汎用的な検索エンジンが構築、提供されている。

また、このような汎用目的の検索エンジンでは対応できない詳細な検索要求に対応するため、特定分野の情報のみを対象にした検索サービスを提供する専門検索エンジンも多数存在している。更に、要求に見合う Web ページを見つけ出すことに留まらず、Web 上に存在する情報に対して、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[5]。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、対象とする特定分野（トピック）のデータを、高品質かつ多数収集しなければならない。そのため、特定トピックのページを効率良く収集するシステムが必要である。

我々は、トピック判定とリンク選出戦略機能を連携させる”One man & His dog”と呼ぶシステムを搭載した自己学習型トピッククローラー（G-CRAWLER）を試作した。本論文では G-CRAWLER について説明し、これを用いた収集実験及び性能評価について述べる。

本論文の構成は以下の通りである。第２章では一般的な Web クローラーについて述べる。また、一般的な収集方法と特定分野（トピック）の収集の違いについても述べる。第３章では関連研究について簡単に紹介する。第４章では開発した自己学習型トピッククローラー（G-CRAWLER）を紹介し、その特徴的なシステムである One man & His dog システムについて述べる。第５章では２つのトピックについての収集実験を述べる。第６章では実験の結果についての議論と考察を述べる。最後に第７章でまとめと今後の課題を述べる。

2. Web クローラーと Web ページの収集

一般的なクローラーについて簡単に説明する。また、一般的な収集方法と、特定分野（トピック）の収集の違いについても述べる。

2.1. クローラーとは

Web ページの収集を人手で行うには莫大な時間と人件費がかかる。そこで、人に代わって Web ページの自動収集を行うプログラムが必要になる。このためのプログラムが Web クローラーと呼ばれるものである。Web クローラーは、『Web ロボット』、『Web スパイダー』と呼ばれることもある[9]。Google などの検索エンジンでもこの Web クローラーを用いている。

2.2. クローラーの Web ページ収集のしくみ

Web クローラーが Web ページを取得していく仕組みを示す。

まず与えられた URL 集合から次に取得する URL を選ぶ。そして Web サーバにアクセスし、収集可能なら Web ページをダウンロードし、保存する。取得したページから、リンクの抽出を行い、URL 集合に追加する。以下、この処理を繰り返すことで、クローラーは Web ページを収集していく。処理の流れ図を図１に示す。

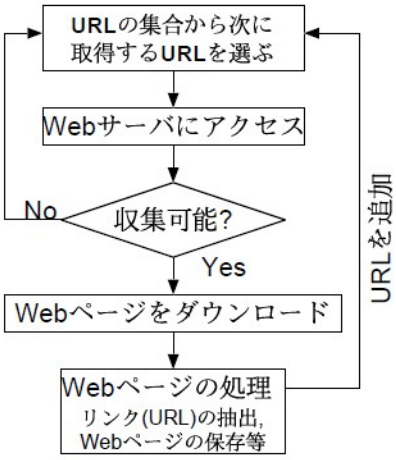

図 1　Web クローラーの処理の流れ図

2.3. 一般的な収集と特定分野の収集

一般的な収集方法では、幅優先探索を用いることが多い。この収集方法では、Web ページを広く、浅く収集していくことになる。指定した範囲のページをもれなく収集できる、プログラムへの実装が容易であるなどの利点がある。

『レシピのデータが欲しい』、『求人情報のデータが欲しい』などのある特定分野（トピック）について収集する場合には、幅優先探索では効率の良い収集ができない。ここで、効率の良い収集とは、『できるだけ必要な Web ページだけを集め、必要の無いページはできるだけ集めない収集法』とする。特定のトピックに関するページを効率よく集めるためには、何かうまい手法を考える必要がある。

一般的な収集と、特定分野の収集のイメージの違いを図 2 に示した。

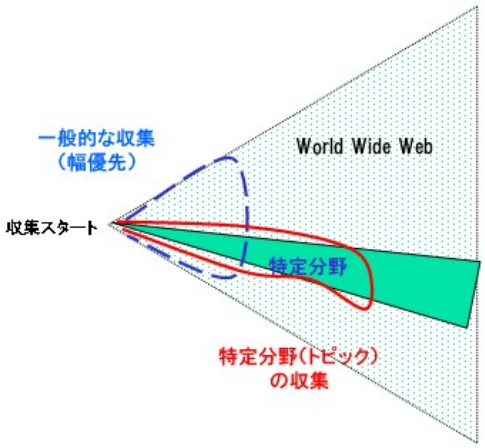

図 2　一般的な収集と特定分野の収集

3. 関連研究

Chakrabarti[7,8]らは、収集したいトピック X に関連するページから収集を始め、X に関連する Web ページを選択的に収集していく Focused Crawler について研究している。例えば『レシピ』を集めたいという要求があったとしよう。レシピが Web 上にどのように存在しているかと考えると、料理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、『似ているページは Web 上において近い場所に存在している』⇒『ターゲットに似ているページ（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はずだ』というのが、このクローラーの発想である。ちなみに、トピック X への関連度は、事前によく学習された文書分類器で計算する。

Diligenti[6]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 3 に context graph の概念図を示す。

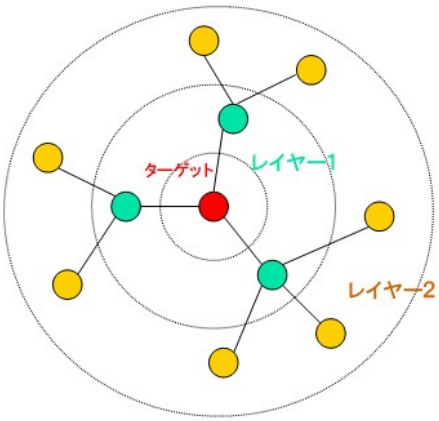

図 3　context graphs の例

ターゲットに 1 回のリンクで辿り着くページ群をレイヤー1、2 回のリンクでたどり着くページ群をレイヤー2 とし、以下、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る。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。

Martin 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。

4. 自己学習型トピッククローラー

我々は G-CRAWLER と呼ぶ自己学習型トピッククロ

ーラーを開発している[10]。G-CRAWLER の概要と、内部に搭載している One man & His dog システムについて述べる。

4.1. G-CRAWLER の概要

図 4 に G-CRAWLER の処理の流れを示す。

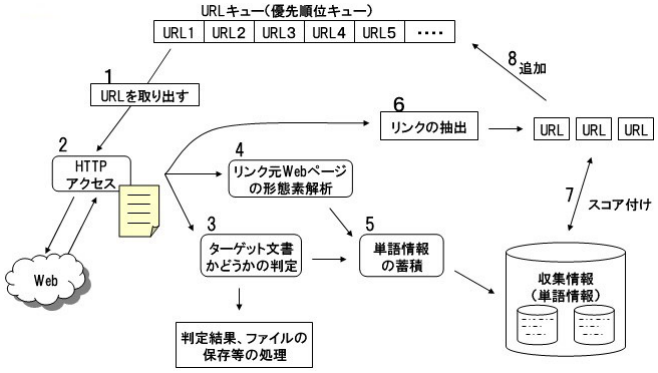

図 4　G-CRAWLER の処理の流れ

G-CRAWLER の特徴は、リンク先ページがターゲットであるかどうかを判定する関数と、リンク元ページの評価する判定関数との、二つの判定関数を持っていることである。さらに、リンク元ページ判定関数は、学習機能を持っている。このシステムを、One man & His dog システムと名づけている。次節で One man & His dog システムについて説明する。

4.2. One man & His dog システム

G-CRAWLER が搭載する One man & His dog システムはリンク先ページがターゲットかどうかを判定する関数を犬（dog）に見立て、リンク元ページのスコアを判定する関数を人間（man）に見立てるシステムである(図 5 参照)。リンク元ページを判定する「人間」は、学習を行う。

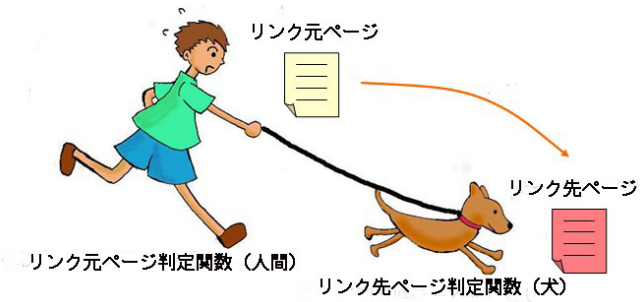

図 5　One man & His dog システムのイメージ

「人間」の学習機能について述べる。犬がリンク先ページをターゲットであると判定した場合、人間はリンク元ページを形態素解析して、単語情報を抽出し、正例としてデータベースに登録する。つまり、『ターゲットに結びつくページ出現する単語』を学習する（図 6 参照）。

犬がリンク先ページをターゲットではないと判定した場合も同様の処理を行う。この場合は、負例としてデータベースに登録する。つまり、『ターゲットに結びつかないページに出現する単語』を学習する。

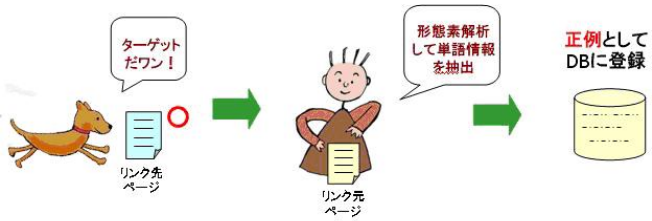

図 6　犬がターゲットを発見した場合

正例、負例の単語情報はそれぞれ以下のような形で足し合わせていく。

$$P_{t+1}(w_i) = P_t(w_i) + \sum P(w_i)$$

$$N_{t+1}(w_i) = N_t(w_i) + \sum N(w_i)$$

ここで、w_i (i=0,1, ...) はページに出現する単語、P は正例の単語情報（Positive Word）を格納した配列、N は負例の単語情報（Negative Word）を格納した配列である。また、tはクローラーが辿るページの訪問順番である。

例えば、3000 番目のページ(t=3000)までで、$P(w_i)$の値が 45 だとする。3001 ページ目に単語w_iが 5 回出現し、かつそのページがターゲットとなるトピックのページであるとする。この場合、$P(w_i)$の値を+5、つまり 50 にする。ターゲットではない場合、$N(w_i)$の値を＋５する。

人間は次に辿るページを決めるために、収集ページから URL を抜き出す。その際、前述の P と N をもとに URL へ次の式でスコア付けをする。

$$\mathrm{Score}(URL) = \sum \frac{P_t}{P_t + N_t}$$

正例の単語を多く含むページからのリンクはスコアを高く、負例の単語を多く含むページからのリンクはスコアが低くなる。そして、スコア付けされた URL を URL キューに追加する。図 7 にスコア付けの様子を示す。

One man & His dog システムは、ターゲットへのリンクが張られていそうなページからのリンクを優先するという、単純な考えに基づいている。人間が学習し

た知識に基づいて次に訪れる場所を決定し、犬をその場所に行かせる。そして、犬が訪れたページを判定し、その結果によりまた人間が学習する・・・といったイメージである。

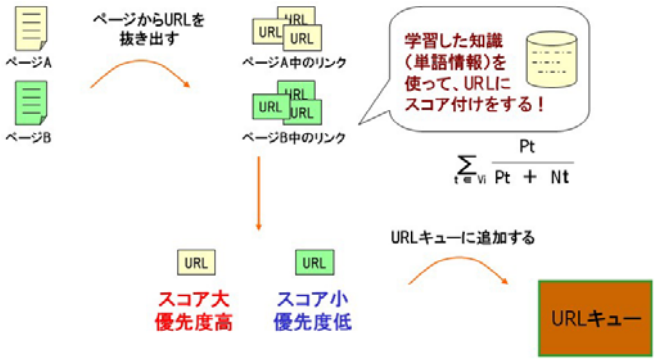

図 7 URL のスコア付け

4.3. リンク集とターゲット

One man & His dog システムで効率よく集まるページ郡の構造について述べる。Web において、ある特定トピックに関するページを集めたリンク集ページを作ることは、よく行われている。

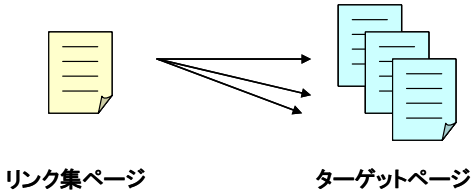

図 8　リンク集とターゲットの構造

すなわち、図 8 のようなリンク集ページが見つかれば、大量のターゲットが得られる。リンク集ページの特徴を学習し、そこからのリンク先を優先して収集することで、ターゲットページの収集効率を上げるのが One man & His dog システムの発想である。

5. 収集実験

この章では、G-CRAWLER の実験と性能評価について述べる。まず、実験の評価指標および 3 つの実験方法について述べる。次にレシピの収集実験およびニュース記事の収集実験について述べる。

5.1. 評価指標

G-CRAWLER を評価するに当たって、次の 2 つの指標を考える。

① ターゲット判定関数の評価
② 収集速度
③ 学習機能の評価

①は収集したページが、集めたいトピックのページであるかどうかの正確さである。つまり、One man & His dog システムにおいては、どのくらい正確に判断できる犬なのかということになる。この評価には、適合率（precision）を使う。

$$precision = \frac{\text{真のターゲット数}}{\text{システムが見つけたターゲット数}}$$

②の収集速度は、トピックのページへの辿りやすさを表す。訪問ページ数に対し、取得ターゲット数が多ければ、早い（効率のよい）クローラーであるといえる。この評価には、縦軸を収集したターゲットページ数、横軸をクローラーの訪問ページ数にしたグラフを描く。One man & His dog システムでは、人間のリンク先選択戦略機能の効率をあらわす。

③については、One man & His dog システムにおける人間の学習機能の効率をあらわす。そのために、単なる幅優先探索と、学習機能を有効にした場合とを比べて、収集効率の変化を調査する。さらに、一回目に学習した単語情報を使って再度収集し、学習結果を活用した場合についても収集効率を調べた。

5.2. 実験方法

収集したトピックについて述べる。実験では、「料理のレシピ」（以下レシピ）と「ワールドカップサッカーのニュース記事」（以下、ニュース記事）2 つのトピックを選んだ。

レシピは、トピックかどうかの判定がしやすく、かつ図 8 にある構造を持つ例として選んでいる。「レシピ」という単語を含む文書は、料理レシピ以外にはほとんど存在しないため、判定が容易である。一方、ワールドカップについては、関連するページが多いため、簡単にトピック判定ができない。

次に、収集方法について述べる。実験で用いた表 1 に実験方法を述べる。方法として、ABC の 3 通りの方法を比較した。

方法	内容
A	幅優先探索
B	学習機能 ON
C	学習機能 ON 2 回目

表 1　収集方法

なお、Precision の値は、実験によりクローラーが見つけたターゲットページが、本当にトピックページであるかを人手で一つ一つ判断し、算出している。

5.3. レシピ収集実験

まず、レシピの収集実験について述べる。『味の素レシピ大百科』[3]のサイトから、レシピの収集実験を行った。スタート URL をサイトのトップページに設定した。たどるリンクをサイト内に限定し、他のサイトに飛ばないようにした。そして、1 万ページを収集した。結果を表 2 に示す。

方法	収集数	正解数	Precision
A	3,860	3,860	1
B	6,435	6,435	1
C	6,807	6,807	1

表 2　レシピ集実験　結果

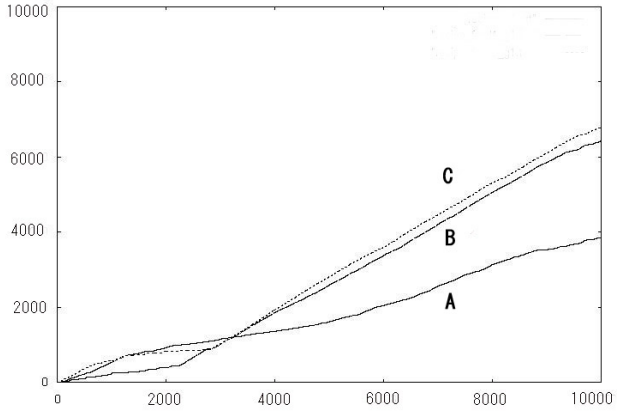

図 9　レシピ集実験

precision が 1 と、最高の値を出しているが、これは、サイト内のレシピは全て、あるテンプレートに沿って書かれており、そのおかげで正確な判定関数を作成することができたからである。レシピページの例は、図 10 に示した。

このように、ターゲットページがある基準にそって書かれていれば、正確なターゲット判定関数を作成することができる。しかし、このサイト以外の、つまり Web 上に存在する他の多くのレシピについては、このサイトのテンプレートに沿って書かれているわけではないので、ここで用いた判定関数が、どんなレシピでも正確に判定できるわけではない。

図 10　レシピページの例

5. 4. ニュース記事収集実験

ここではニュース記事の収集実験について述べる。集めるターゲットは、サッカーW 杯に関するニュース記事とした。ニュースサイト『asahi.com』[4]のスポーツニュースのトップページをスタート URL に指定して、最大訪問ページ数を 1000 ページとした。ターゲットの判定は、出現単語について、

（ワールドカップ∨W 杯）∧ サッカー

を満たすものとした。収集したターゲットに対して、人目で確認を行った結果、次のようになった。

方法	収集数	正解数	Precision
A	11	10	0.909
B	89	87	0.978
C	106	103	0.972

表 3　ニュース収集実験　結果

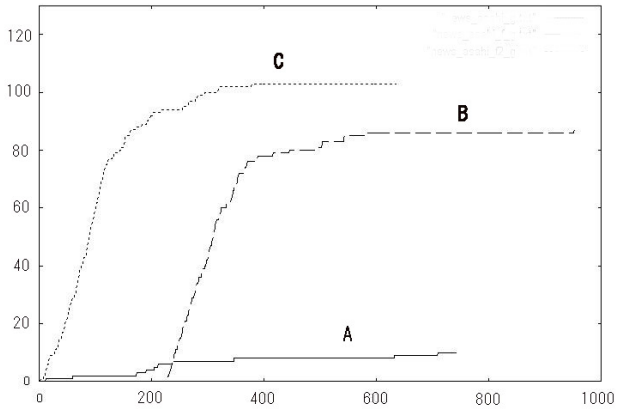

図 11　ニュース収集実験

収集したニュース記事の例は図 12 に示す。

図 12　収集した記事の例

6. 考察

6.1. 学習効果

学習機能の効果について考察する。図 9、図１１からもわかるように、A の幅優先探索よりも B の学習機能を使った場合の方が、明らかに効率が良いことがわかる。

レシピの場合、訪問ページが 1 万ページの時点で、B の学習機能を ON にした場合では、A の幅優先探索の場合に比べて、約 1.67 倍のレシピを収集できている。ニュース記事の場合、1000 ページの時点で B の学習機能を ON にした方が A の幅優先探索の 8.7 倍のページを収集できている。

また、C の蓄積された単語情報を最初から生かした場合は、両方の実験において B よりも効率が上がっている。つまり、学習機能を使った場合（B と C）は、幅優先探索の場合と比べて、収集のスピードが早いということが言える。

ただし、図９、図１１をみると、B の学習機能を ON にした場合では、開始直後は A の幅優先探索よりも効率よく収集できていない。しかし、単語情報が集まってくる中盤以降は、蓄積した知識を生かして、収集効率が大幅に上がっている。つまり、One man & His dog システムにおいては、学習を重ねることで、人間の頭が良くなっていくというイメージである。実験数が少ないので、たまたまこのような結果が出たのかもしれないが、未学習の場合が幅優先探索よりも効率が悪い理由については、今後調査していく。

C の蓄積された単語情報を最初から用いた場合では、スタート直後からでも効率よく収集できていることが分かる。蓄積された知識（単語情報）を次回の収集に生かせるという利点が、One man & His dog システムにはあると言える。

6.2. リンク集とターゲット

レシピや、ニュース記事が 4.3 節で述べたような、リンク集とターゲットの構造をしているかどうかを調べた。その結果、図１３や図１４のようなリンク集ページが見つかった。これにより、リンク集とターゲットの構造をもつページ群に対して、One man & His dog システムは効率良くトピックページの収集を行えことが分かった。

図 13　レシピへのリンク集ページ

図 14　ニュース記事へのリンク集ページ

7. まとめと今後の課題

Web 上の情報量の爆発的増加に伴い、専門検索エンジンの構築や、Web マイニングなどといった研究が盛んに行われている。また、日常の様々な場面で情報技術の利用が進められている。本論分では特定のトピックに関する Web ページを効率よく収集する自己学習型トピッククローラー（G-CRAWLER）について述べた。G-CRAWLER の特徴として、リンク先とリンク元、二つの判定関数をもちいた学習システムである One man & His dog システムがあげられる。この学習システムの効率を評価するために、レシピとニュースについて収集実験を行い、ページがリンク集とターゲットの構造を持っていれば、一般的な幅優先探索の手法よりもトピックページを効率よく集めることができることを示した。

今後の課題としては、まず他の論文で提案されている手法との比較を行う。関連研究の章でとりあげた手法と、我々の One man & His dog システムとの比較実験を行う予定である。ただし、比較のためには他者の収集手法の実装が必要になる。リンク集とターゲット群との構造になっていないトピックでの、One man & His dog システムの効果も確かめる予定である。

最終的には扱いやすいインターフェースの開発、新たな機能などの追加を行い、Web 上からのデータ収集のための強力なツールの開発を目指したい。

文　献

[1] Google : http://www.google.co.jp/

[2] Yahoo! : http://www.yahoo.co.jp/

[3] 味の素レシピ大百科 : http://www.ajinomoto.co.jp/recipe/

[4] asahi.com : http://www.asahi.com/

[5] 鈴木英之進、他, “特集 最新！データマイニング手法”,情報処理, vol.46 No.1, pp.4-51, 2005.

[6] 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.

[7] 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.

[8] 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.

[9] Kevin Hemenway, Tara Calishain, 村上雅章（訳）, “Spidering Hacks”, オライリー・ジャパン, 2004.

[10] Yoshihiro Matsunaga, Shintaro Yamada, Eisuke Ito, Sachio Hirokawa, “A Web Syllabus Crawler and its Efficiency Evaluation”, Proc. of International Symposium on Information Science and Electrical Engineering 2003, pp.565-568, 2003.

[11] 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.