九州大学学術情報リポジトリ
Kyushu University Institutional Repository

自己学習型トピッククローラーの性能評価に関する考察

冨山，北斗
九州大学大学院システム情報科学府

伊東，栄典
九州大学情報基盤センター

廣川，佐千男
九州大学情報基盤センター

http://hdl.handle.net/2324/17787

出版情報：情報処理学会全国大会講演論文集. 69 (1), pp.601-602, 2007-03. 情報処理学会
バージョン：
権利関係：

KYUSHU UNIVERSITY

自己学習型トピッククローラーの性能評価に関する考察

冨山 北斗 † 廣川 佐千男 †† 伊東 栄典 ††
† 九州大学システム情報科学研究府情報理学専攻 †† 九州大学情報基盤センター

1 はじめに

Web 上には、個人が必要とする量をはるかに上回る情報が存在する。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が、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。我々は、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、トピックページへ早く辿りつ速度が重要である。我々は、One man & His dog 方式と呼ぶトピック判定とリンク選出戦略機能を連携させ、自己学習を行いながら収集する方式に基づくクローラー (G-CRAWLER) を試作し、既存の収集方法と収集速度について比較した。また、学習の有効性についても調査した。本論文ではこの比較調査について述べる。

2 自己学習型トピッククローラー

Web 上において、あるトピックに関連するページは、図 1 のように、一つのリンク集のようなページからまとめてリンクが貼られていることが多い。

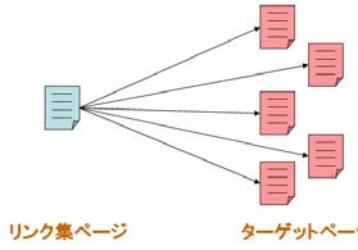

図 1: リンク集とターゲット

G-CRAWLER は、リンク先ページがターゲットか否かを判定する関数と、リンク元ページを評価する関数との、二つの判定関数を持っている (図 2)。この方式を、One man & His dog システム (以下、OHS) と名付けている。リンク先ページがターゲット否かを判定する関数を “犬”、リンク元ページのスコアを評価する関数を “人間” に見立てている (図 3)。

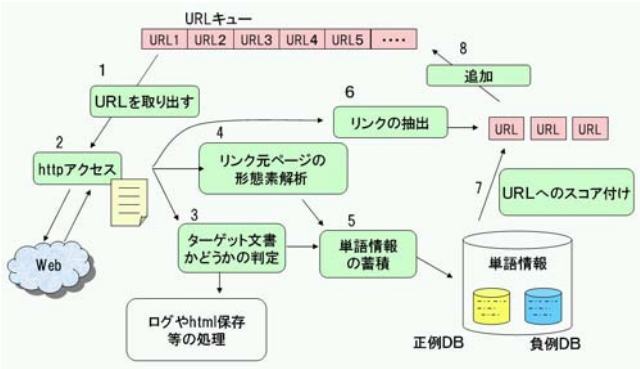

図 2: G-CRAWLER の処理の流れ

図 3: One man & His dog 方式のイメージ

“人間” は犬の反応を見てリンク集ページの特徴を学習し、その特徴に似たページからのリンクを優先して収集するよう、犬を誘導する。G-CRAWLER 及び One man & His dog 方式についての詳細は文献 [2, 4] で述べている。

3 速度比較実験

OHS の有効性を調査するために、他の手法と収集速度の比較実験を行った。

3.1 比較手法

[a〜d:学習/事前知識 有り無し]

OHS では前回クロールした際の学習結果 (知識) を次回の収集で利用可能である。OHS の設定として事前知識を用いるか否か、また、収集中学習を行うか否かで 4 通りに分類可能である。事前知識を用いず学習も行わない場合は幅優先探索を用いた。(表 1 参照)

Perfomance evaluation of self-adaptive topic focused crawler

† Hokuto TOMIYAMA (hokuto.tomiyama@i.kyushu-u.ac.jp)
†† Sachio HIROKAWA (hirokawa@cc.kyushu-u.ac.jp)
Eisuke ITO (itou@cc.kyushu-u.ac.jp)
Dept. of Informatics, ISEE, Kyushu Univ. (†)
Computing and Communications Center, Kyushu Univ. (††)

表 1: OHS について 4 通り

	事前知識有り	事前知識無し
学習 ON	a:OHS1	b:OHS3
学習 OFF	c:OHS2	d:幅優先

[e:類似性]

ターゲットの特徴に近いページからのリンクを優先する手法である。この手法は多くのトピッククローラーで利用されている ([1])。

[f:アンカーテキスト]

リンクの優先度決定にアンカーテキストを使用する手法である。この手法も様々なトピッククローラーで利用されている ([3])。

[g:ランダム]

リンクの優先度をランダムに決定する手法である。他の手法との比較のために実装した。

3.2 実験

縦軸を収集ターゲット数、横軸を訪問ページ数として収集速度の比較を行った。

レシピ収集実験

『味の素レシピ大百科』のサイトより、レシピページの収集実験を行った（図 4)。レシピ収集では、OHS がもっとも収集速度が速かった。サイト内を調査したところ、図 1 のようなリンク集とターゲットの構造をしていることが分かり、この構造への、OHS の有効性を示すことができた。また事前知識を用いれば、さらに収集速度を上げることができた。

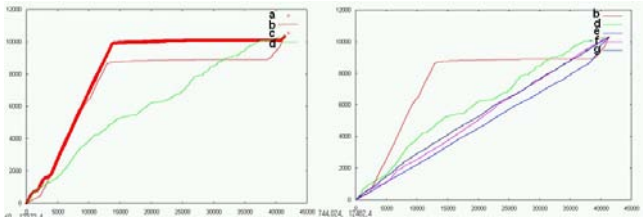

図 4: レシピ収集実験

募集要項収集実験

企業 100 社のサイトから、募集要項のページを収集した（図 5)。募集要項はサイト内に高々数ページしかなく、リンク集とターゲットの構造をしていない。しかし、募集要項にたどり着くページの特徴を “人間” がうまく学習できているので、OHS の収集速度が速くなっている。また、事前知識を用いた場合はさらに収集速度を上げることができた。これにより、“人間” が学習することにより得られる知識の有用性が分かる。

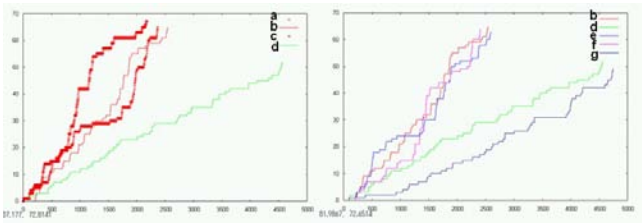

図 5: 募集要項収集実験

決算報告収集実験

募集要項と同様の 100 社のサイトから決算ページを収集した (図 6)。アンカーテキストと似たもの同士手法が収集速度が速い。OHS はその 2 つよりは劣るものの、幅優先探索よりは収集速度が速い。

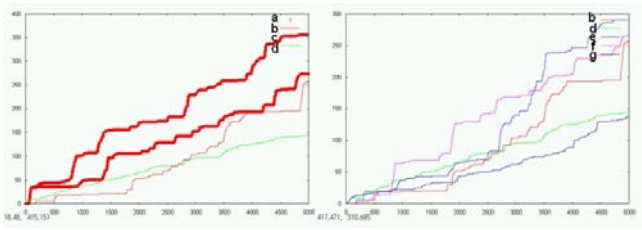

図 6: 決算報告収集実験

4 まとめ

トピック判定とリンク選出戦略機能を連携させ、自己学習を行いながら収集する One man & His dog 方式を考察し、それに基づくクローラーを実装した。他の手法との収集速度の比較により有効性を検証した。レシピ収集実験のように、リンク集とターゲットという構造を持つ場合に非常に有効であった。また、募集要項収集実験のように、ターゲットがごく少数しか存在しない場合でも有効に働く場合があるという結果が得られた。また、事前知識による収集速度向上が確認できた。

参考文献

[1] Soumen Chakrabarti et al., ”Focused crawling: a new approach to topic specific Web resource discovery”, Computer Networks, Vol.31, No.11-16, pp.1623-1640, 1999.

[2] 廣川佐千男, 松永吉広, 野口正人, ウェブ情報収集装置とウェブクローラープログラム、及びウェブ情報収集方法, 特願 2004-168034.

[3] Jun Li, Kazutaka Furuse, Kazunori Yamaguchi, ”Focused Crawling by Exploiting Anchor Text Using Decision Tree”, WWW 2005, May 10.14, 2005.

[4] 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, 3B-i11, March.2, 2006.