E-017

子供向け Web サイト収集のためのクローリング手法の検討
Consideration of a Crawling Method for Collecting Web Sites for Children

泉川 洸一郎† 安藤 一秋‡
Koichiro Izumikawa Kazuaki Ando

1. はじめに

近年，小学校などの教育機関では，新聞を教材として活用する教育（NIE : Newspaper In Education）が実施されている． NIE の実態調査結果報告書[1]によると，NIE を実践することで，児童の読解力と表現力が向上することや，社会に対する関心が高まることなどが報告されている. しかし，新聞に出現する語句は, 小学生にとって難しいものが多く，記事の理解は容易ではない.

この問題を解決するため，新聞記事に出現する難しい語句を小学生が理解できる平易な語句に言い換える研究[2]が進められている．難しい語句を平易に言い換えるためには，言い換え知識が必要となる．小学生を対象とした言い換え知識として小学国語辞典が利用できるが語彙数が少ない．したがって，語彙数問題を解決するためには，言い換え知識の自動獲得が必要となる.

近年，大規模なコーパスが整備されており，様々な用途に活用できるようになった．しかし，子供向けに書かれたテキストを大量に集めたコーパスは, 現在存在していない.

本研究では，Web 上の子供向けサイトに存在する平易なテキストを大量に収集することで「子供 Web コーパス」を構築し，コーパスから言い換え知識を自動獲得する手法の実現を目的とする．本稿では，事前調査として，クローリングにより子供向けサイトが収集できる可能性を調査する.

2. 子供向け換言に関する関連研究

梶原らは，学習基本語彙ではない語を難解語と定義し，難解語を学習基本語彙に言い換える手法[2]を提案している．複数の国語辞典から難解語を見出し語として検索し，定義文から難解語と同一品詞の学習基本語彙を換言先候補として取り出す．そして，取り出した換言先候補を語の類似度など複数の指標を用いて言い換える．梶原らは言い換え知識を取得する国語辞典の一つとして小学国語辞典を使用しており，問題点として収録されている語彙数が少ないと指摘している．本稿では，Web 上の子供向けサイトから言い換え知識を収集することで, この問題の解決を目指す.

3. 子供向けサイトのクローリング手法の検討

Web 上から子供向けに書かれたテキストを収集するため, 子供向け Web ページ内の外部リンクに注目する. 子供向けページ内にある外部リンク先は，他の子供向けサイトである可能性が高いと仮定する．本稿では，子供向けサイトをシードとしてクローリングし，大量の子供向けサイトとサイト内の Web ページが収集できる可能性を調査する.

子供向けサイトからリンクされている外部サイトは，すべて子供向けであるとは限らない．そこで，子供向けサイトの収集精度を向上するために，小島らが開発した難易度推定システム「帯 2」[3]と子供向けサイトに含まれる特徴的なキーワードを用いてノイズの削減を試みる.

3.1 帯 2 による難易度の推定

帯 2 は，各学年別の教科書コーパスを活用することで，与えられたテキストの難易度を小学 1 年生から大学生以上までの 13 段階で推定する[3].

事前調査として，子供向け Web ページ 296 件に対して，帯 2 を用いて難易度を推定した結果，難易度が小学 1 年生から中学 1 年生までの間に推定されたページが 282 件（95.3%）という結果が得られた．したがって，Web ページの難易度が中学 1 年生以下と推定された場合，子供向け Web ページとして収集できる可能性がある.

3.2 特徴的なキーワードによる推定

子供向けサイトには，訪問者に子供向けであることを示すために，各ページに「こども向け」や「キッズページ」など，子供に関連するキーワードが示されることが多い. また，大人向けページと子供向けページを区別して管理するために，URL に「kids」などのキーワードを含んでいる場合もある．クローリング時に，これらのキーワードがページ内や URL に含まれていれば，子供向け Web ページとみなして収集できる可能性がある.

4. 調査

子供向けサイトをシードとしてクローリングし，他の子供向けサイトが収集できる可能性について調査する. また, 帯 2 と特徴的なキーワードによる子供向けサイトの判定性能を調査する.

4.1 シードの選択

Yahoo!きっずは, 子供を対象とした検索サービスであり, 人手により子供向けと判断されたサイトを紹介している. 本調査では，Yahoo!きっずで紹介されている外部サイトへのリンクが豊富な「北陸農政局キッズページ[4]」を選択し，シードとして利用する.

4.2 子供向けサイトの収集調査

子供向けサイトをシードとして，簡易クローリングした結果，80 件の外部リンクが収集できた． 80 件の外部リンクを人手で確認した結果，その中には子供向けサイトが 15.0%（12/80 件）含まれていた．この 12 件には，シードに利用したサイトとは異なる子供向けサイトが存在していた．したがって，新たに収集した子供向けサイトから再帰的にクローリングすることで，大量の子供向けサイトを収集できる可能性がある.

しかし，クローリングした 85.0%（68/80 件）は，子供 Web コーパスに利用できない一般のサイトであった．したがって，クローリングによって大量の子供向けサイトを収集するためには，ノイズを削減する必要がある.

4.3 収集精度の向上調査

帯 2 と子供向けサイトに特徴的なキーワードを利用して, 子供向けサイトの判定性能を調査する. 利用するデータは, 4.2 節で収集した 80 件を利用する.

† 香川大学大学院工学研究科 Graduate School of Engineering, Kagawa University
‡ 香川大学工学部 Faculty of Engineering, Kagawa University

まず，帯 2 の難易度判定を利用する方法は，リンク先の Web ページの難易度が中学 1 年生以下と推定された場合，リンク先のページおよびそのページからリンクされている階層の深いすべての内部サイトの Web ページを子供向けとして収集する．一方，特徴的なキーワードで判定する方法は，リンク先の URL に「kids」が含まれているか，リンク先のページ内に「こども」「キッズ」「小学生」など特徴的なキーワードが含まれている場合，そのページおよびそこからリンクされている階層の深いすべての内部サイトの Web ページを収集する．

2 つの手法で収集した結果を表 1 に示す．表 1 に示すように，キーワードによる収集法による F 値は 0.48 と判定手法なしと比べて性能が向上していることから，ノイズサイトの収集を減らすことができたといえる．しかし，精度は 46.2%と低く，再現率も 50.0%であるため，子供向けサイトの抽出漏れが増加している．一方，帯 2 の難易度判定による収集は，精度と再現率が共に低い．したがって，今後は手法の改良や新しい手法の検討が必要である．

表 1　子供向けサイトの収集結果

判定手法	総抽出件数	子供向けサイトの抽出件数	精度	再現率	F 値
なし	80	12	15.0	100.0	0.26
帯 2	39	4	10.3	33.3	0.16
キーワード	13	6	46.2	50.0	0.48

5. 考察

5.1　収集できなかった子供向けサイトの特徴

4.2 節で収集した 12 件の子供向けサイトの内，帯 2 やキーワードによる手法で収集できなかったサイトの特徴を表 2 に示す．どちらかの手法で収集できない子供向けサイトは合計 9 件あり，4 件が画像や Flash のみでページが構成されていた．その内の 2 件は，URL により子供向けと判定できるが，残りの 2 件は最初のページのみで判定できない．これらのサイトは，リンク先のページからさらにリンクされている内部ページを判定に利用することで，収集できる可能性がある．

また，両手法でも収集できない子供向けサイトは，5 件あった．その中には文字コードが自動判別できないサイトが 2 件含まれていた．文字コードが判別できない場合，帯 2 もキーワードによる手法も適用できない．

5.2　子供向けサイトに含まれるノイズの削減

子供向けサイトを構成しているページの多くは子供が理解しやすい文で記述されているが，保護者や先生を対象としたページも含まれている．これらの大人向けページを帯 2 で判定できる可能性を調査する．

具体的には，4.2 節で収集した 12 件の子供向けサイトから大人向けページを含むサイトを 2 件選択し，サイトを構成する各ページの難易度を帯 2 で推定することで，子供向けか否かで分類する．そして，2 種類に分類されたページから各 10 件をランダムで選択し，その分類結果の妥当性を人手で確認する．

帯 2 による子供向けページの分類結果を表 3 に示す．結果を確認した結果，帯 2 の難易度推定によって子供向けと分類された 20 件のページには，大人向けと誤って判定されたページは存在しなかった．しかし，本調査で対象としたページ数が少ないため，今後は対象ページ数を増やして調査する必要がある．

帯 2 による判定で，大人向けに分類された Web ページの内，サイト 1 の 5 件とサイト 2 の 3 件は，子供向けページであった．小島ら[3]の調査によると，小学生向けと明示されている Web ページの難易度の平均値は中学 1 年生と推定されている．本調査では，帯 2 による難易度判定結果が中学 1 年生までを子供向けページとして分類しているため，中学 2 年生以上と推定された子供向けページが大人向けに分類されたと考える．

表 2　収集できなかった子供向けサイトの特徴

サイトの特徴	件数	収集結果	
		帯 2	キーワード
(1) トップページが画像や Flash のみで構成	2	×	×
(2) 文字コードが自動的に判断できない	2	×	×
(3) 難易度が高く，キーワードが含まれない	1	×	×
(4) URL に「kids」が含まれかつ特徴(1)のサイト	2	×	○
(5) 難易度が高いがキーワードが含まれる	1	×	○
(6) 難易度が低いがキーワードは含まれない	1	○	×

表 3　難易度推定によるページの分類結果

	適切な分類件数	
	子供向け	大人向け
サイト 1	10	5
サイト 2	10	7

6. おわりに

本稿では，子供向けサイトをシードとしてクローリングすることで，他の子供向けサイトが収集できる可能性を示した．しかし，単純なクローリングではノイズが多く，帯 2 や子供向けサイトに特徴的なキーワードを利用しても収集精度は 50%程度であることを確認した．

今後は，これらの手法を改良すると共に機械学習による判定などを検討し，子供向けサイトの収集精度の向上と子供 Web コーパスの構築を目指す．

謝辞

本研究の一部は，JSPS 科研費 25350335 の助成を受けて実施した

参考文献

[1] NIE 実践の実態調査結果報告 http://nie.jp/inves/ji1_200807.pdf

[2] 梶原 智之，山本 和英，“語釈文を用いた小学生のための語彙平易化”, 情報処理学会論文誌, Vol56, No.3, pp. 983-992, (2015).

[3] 小島 健輔，佐藤 理史，藤田 篤，“文字 bigram モデルを用いた日本語テキストの難易度推定”，言語処理学会第 15 回年次大会論文集，pp. 897-900，(2009).

[4] 北陸農政局キッズページ http://www.maff.go.jp/hokuriku/kids/