

西埜 覚\* 西塚 一樹\* 苗村 憲司\*\*

通信・放送機構 横浜コンテンツリサーチセンター\* 慶應義塾大学環境情報学部\*\*

## 1. はじめに

小中学校の授業でインターネットを利用する場合、生徒に閲覧させることが授業目的の達成を妨げる情報(以下では有害な情報という)の閲覧を防ぐ必要がある。公開される情報のコンテンツ(ウェブページ)を複数カテゴリ、複数レベルで有害な情報の格付けをする方式(多段階レーティング方式)が考えられている。これを適用するフィルタリングソフトを用いれば、目的にあった情報のみを閲覧することができるので、授業の目的にあった情報の閲覧が可能になる。WWWコンソーシアム標準のPICS(Platform for Internet Content Selection)はこの格付け情報を記述する構文で、これに基づく一方式のRSACi (Recreational Software Advisory Council on the Internet)では、4カテゴリ・5レベルの格付け基準を定めており、国際的な基準作りも推進されている。

格付けの方法には、「自己格付け」と「第三者格付け」がある。第三者格付けは、作業の大半が人手に頼るため、多大な工数を必要としており、その省力化が求められている。

多段階レーティング方式を前提に、ウェブページをテキスト処理で有害な情報の数値化、ページ間のリンクの関連付けによって有害な情報の検出を行い、格付け作業の省力化を実現する技術の研究開発を行なっている[1],[2]。

## 2. 有害な情報におけるページリンクの特徴

図1に示すようなウェブページのリンク関係において、「ページXの作成者はページYの情報」を理解してリンクを設定」しており、ページYが有害な情報を含む場合、それをリンク先にしてページXも有害な情報を含む可能性が高いといえる。例えば、ページYが暴力的な情報を含む場合、それをリンク先に行っているページXも暴力的な情報を含む場合が多い[1],[2]。有害な情報を含むウェブページのリンク関係におい

て、ページXの格付けにリンク先ページYの格付けの結果を利用することが有効であると考えられる。

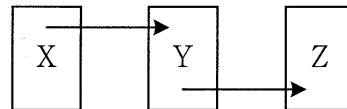


図1 ページの相互リンク

## 3. リンク先情報の先取りの可能性

図2に示すウェブサイトとページの関係において、同一のウェブサイト(ページの作成者が同一の場合が多い)にページが存在(ページX,Y,Z)してリンクを設定している場合、高い割合でリンク先ページの格付け情報を利用できること(以下ではリンク先情報の先取りという)を明らかにした[2]。異なるウェブサイトが存在するページ(ページVとX、ページZとW)の間においても、リンク先情報の先取りが有効であることが示されれば、ウェブページの格付けの支援に効果がある。

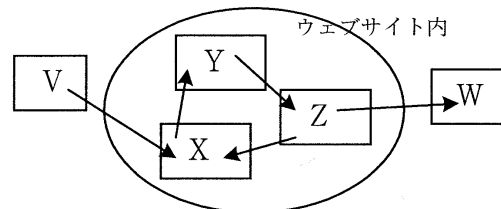


図2 ウェブサイトとページ

## 4. 異なるウェブサイトのページ間でのリンク先情報の先取りの実験

個別のページでテキスト処理を行ない、4カテゴリに分けて有害な情報の数値化を行なった。数値化の後、異なるウェブサイトのページをリンク先に持つものを約3,800ページ抽出して、リンク先ページとの関連性の確認を試みた。

<sup>1</sup>A possibility of using the information forward-linked pages for detecting harmful contents on WWW, Satoru NISHINO, Kaduki NISHIDUKA, Telecommunications Advancement Organization of Japan, and Kenji NAEMURA, Keio University

なお、この3,800ページには、FRAMEタグ等による表示ページの組み込みや、リンク集のように多数のページを持つものが含まれている[3]。また、以降では、1カテゴリ以上で有害な情報の数値化の結果が閾値を越えるものを「有効」、越えないものを「無効」ページという。

- (1) 閾値をゼロとして異なるウェブサイトが存在するページの関連性を調べた結果を図3に示す。グラフWは、リンク先に同一のカテゴリで有効なページを持つもので、全体の65.2% (2,514ページ) を占める。グラフZは、異なるカテゴリで有効なページで3.7% (141ページ)。グラフYは、自己が有効・リンク先が無効なページで12.9% (499ページ)。グラフXは、自己が無効・リンク先が有効なページで18.2% (700ページ)を占めている。
- (2) リンク先が有効なページであるものは、87.1% (グラフX,Z,Wの合計) を占める。これらのページについては、リンク先ページ情報の先取りが有効である可能性を持つ。
- (3) 自己が無効なページ (グラフX) や異なるカテゴリで有効なページ (グラフZ) が検出される場合、リンク先ページ情報の先取りが特に有効である。この2つについて、有害な情報を明確にするために閾値を0.5に定めて実験した結果を図4に示す。自己が無効 (グラフM、700ページ) では56.4%、異なるカテゴリ (グラフN、141ページ) では46.8%のページでリンク先の情報の先取りが可能である。また、この割合はテキスト処理の改善や閾値の設定で向上させることが可能である。
- (4) (3)と同じ閾値の条件で、グラフMの自己ページの文字数とリンク先情報の関連を調べた (図5)。100文字以下では74.8% (グラフP、151ページ中)、200文字以下では36.1% (グラフQ、97ページ中) でリンク先情報が有効である。このことから、文字数が少ない方が情報の先取りの可能性は高いと思われる。

## 5. まとめ

実験の結果から、同一のウェブサイトが存在するページ間と同様に、異なるウェブサイトのページ間においてもリンク先ページの情報の先取りは可能で、コンテンツの格付け作業を支援するために効果があると判断できる。また、テキスト処理によるページの有害な情報の数値化の精度を高めることや、有害な情報と判定するための閾値の設定が重要である。

今後は今までの実験の結果を踏まえて、多段階レーティング方式の格付け支援を行なうプロトタイプシステムを開発し、学校の現場で利用して評価を試みたい。

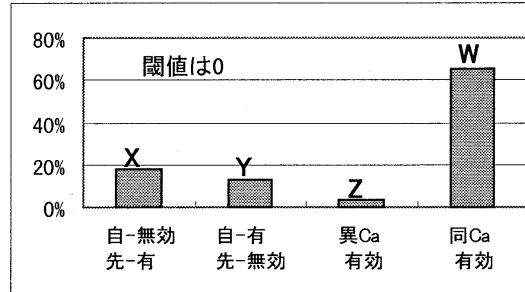


図3 異なるサイトのページとの関連性

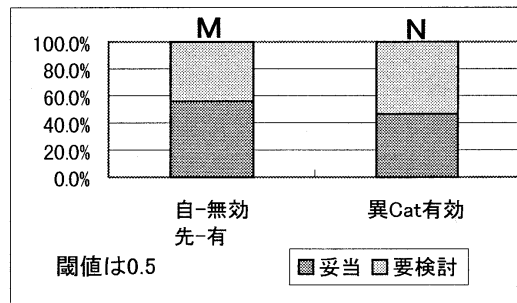


図4 リンク先情報の妥当性

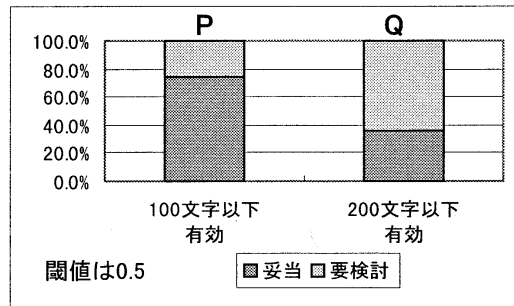


図5 文字数とリンク先情報の関連

## 参考文献

- [1] 西埜、他：“WWW 上の有害な情報を効率的に検出する一手順”，信学会 99 年総合大会,D-15-17(1999)
- [2] 西埜、他：“WWW の教育用格付けの効率化技術に関する検討”，情処学会 99 年後期全国大会,3X-04(1999)
- [3] 西塚、他：“WWW の有害情報検出に関わるある種のリンク機能についての一考察”，情処学会 2000 年前期全国大会,5U-08(2000)