WWWの教育用格付けの効率化技術に関する検討[1]

3X－4

西埜 覚*、苗村 憲司**

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

1．はじめに

小中学校においてインターネットの積極的な活用を進めるにあたっては、生徒に閲覧させることが授業目的の達成を妨げる情報(以下では有害情報という)を遮断することが必要となる。しかし、何が有害情報か、またどの程度有害であるか、の判断は授業目的によって異なり、画一的な区分は不可能である。そこで、それぞれのコンテンツ(ウェブページ)に対して、有害となる可能性のある複数の性質(カテゴリー)の各々について複数レベルの格付けを行っておき、受信側ではこの格付け情報に基づいて適切なレベルでフィルタリングする方式が最も有望と考えられる。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として有用である。PICS に基づく方式の一つである 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進展している。

これらの方式を実用化するためには、ウェブページを格付けしておくことが必要である。ページの提供者が自ら（自己）格付けするにせよ、ISP や教育関係団体などが（第三者）格付けをするにせよ、格付け作業の工数を削減することが重要である。そこで、格付け作業を省力化するための技術を開発することをねらいとし、テキスト処理とハイパーリンクによる接続関係の応用について研究を進めてきた。

2．テキスト処理による格付け支援技術

各ページの記述内容から①キー単語、②2 単語組み合わせ、③文節（短い文）、④自己 URL 中の語句、⑤リンク先ページ URL 中の語句 の出現を検出し、個々の重みと頻度 で重み付けを行い、これを基に有害度の数値化を行うことを試みた。

3．ページのリンク関係を利用した格付け支援技術

一般に、リンク先のページ（図 1 でページXからページYの方向）は容易に参照できるが、逆方向（図 1 でページYからページXの方向）にリンク（逆リンク）をたどって元のページを参照することは不可能でありリンクが設定されたことも不明である。即ち、次の前提を置くことができる。

・ページYの作成者は、ページZの情報を理解しているが、ページXの情報を知らない。

学校教育目的で有害情報の格付けを行う場合、(図 1 を例として) 次のように考える[1]。

・ページYが有害情報として格付けされれば、ページXの格付けにあたってページYの格付け結果を転用することが有効と考えられる（例えば、ページYが暴力的な情報を含む場合、ページXも暴力的な情報を持つ可能性が多い)。

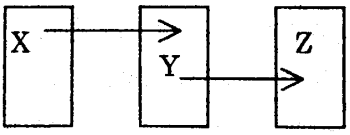

図 1 ページの相互リンク

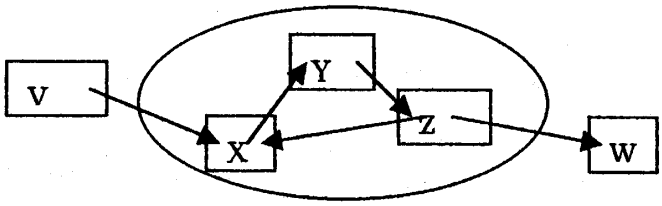

図 2 リンクの認識範囲

[1] A study on technologies for efficient rating of WWW pages for educational purposes,

Satoru Nishino, Telecommunications Advancement Organization of Japan, and Kenji Naemura, Keio University

２と３の技術を組み合わせることにより、効率的な格付け支援技術の実現が期待される。

４．評価結果例

あるプロバイダのサーバにある12個人ユーザの約1530ページを対象とし、次の手順により有害情報の格付け評価を行った。表１、図３、図４にその結果を示す。

(1) 何らかの有害カテゴリについて、テキスト処理で格付けされたページを「有効なページ」とする。

(2) 対象となる範囲(図2)で有効なページから見て先リンク（ページＸからＹの方向)、逆リンク（ページＹからＸの方向）のページが、同一カテゴリで格付け数値を持つ時「リンクに有効有り」とする。

表１　数値化、リンク関連の評価結果

	リンク総数		数値化結果		リンクの関係性				
	a	x	y	f	g=f/a	h	j	k=j/f	m	n
区分	Page数	先Link総数	逆Link総数	有効なPage	有効なPageの割合	先Linkに有効有り	逆Linkに有効有り	逆Linkに有効の割合	先/逆Linkに有効有り	先&逆Linkに有効無し
A	156	2,365	332	106	67.9%	45	71	67.0%	77	79
B	728	7,566	3,377	648	89.0%	539	624	96.3%	629	99
C	151	16,945	3,842	73	48.3%	70	61	83.6%	70	81
D	103	1,360	136	20	19.4%	6	8	40.0%	8	95
E	160	168	160	8	5.0%	0	0	0.0%	0	160
F	111	443	210	62	55.9%	16	36	58.1%	43	68
G	236	1,828	866	92	39.0%	77	78	84.8%	83	153
H	280	2,506	1,037	77	27.5%	41	55	71.4%	66	214
J	101	112	109	80	79.2%	7	79	98.8%	80	21
K	132	517	490	9	6.8%	0	0	0.0%	0	132
L	153	1,110	956	131	85.6%	105	126	96.2%	127	26
M	107	412	264	9	8.4%	0	0	0.0%	0	107

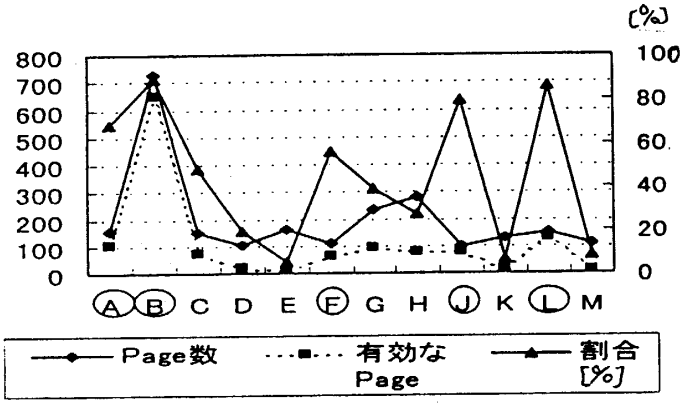

図３　有効なページの割合

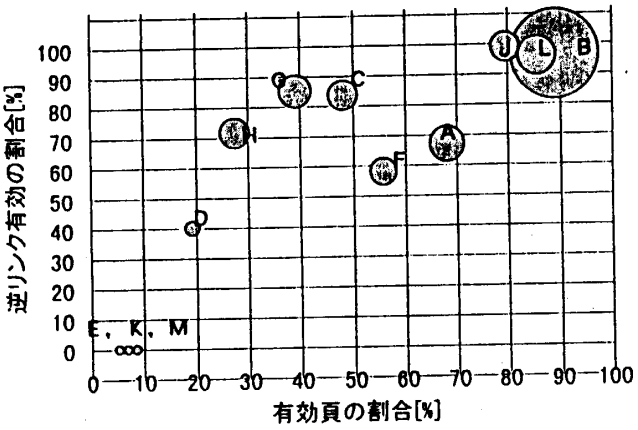

図４　有効なページの割合と逆リンクが有効な割合
(円の大きさは有効なページの数に対応)

評価結果から、上の方式はコンテントの格付けを支援するために有効と判断できる。今後、テキスト処理に用いた辞書データの精緻化と、任意のページにおいて相互リンクの関連性の評価が重要である。

文献

[1] 西埜、満澤、苗村：“WWW 上の有害な情報を効率的に検出する一手順”,電子情報通信学会 1999年総合大会論文集,情報・システム１,P.278(1999)