3X−05

ページ内リンクリスト情報を利用したWebページアクセス制御用URLリスト拡張手法

牧野 俊朗　杉崎 正之　稲垣 博人

日本電信電話株式会社
NTTサイバーソリューション研究所

1 はじめに

インターネット上には、多種多様な情報があるが、子供がインターネットを使用する場合、親が子供に見せることが望ましくないと思うコンテンツに子供がアクセスしてしまうことがあり、このようなアクセスを規制したいという要望がある。また、企業によっては、社員の生産性の低下を防ぐために、業務に無関係なサイトへのアクセスを規制する場合がある。このようなアクセス規制を実現するために、様々なフィルタリングソフトが提供されている。フィルタリングソフトは、基本的に人手で作成したＵＲＬリストを元に、アクセスの可否を決定している。インターネット上では、新サイトの出現および既存サイトの消滅が頻繁の行われているため、このＵＲＬリストの作成およびメンテナンスは多大な労力を必要とする。一方、インターネット上には、様々なリンク集が存在し、ページ作成者がメンテナンスをしている。本稿では、このリンク集を利用することによって、Webページへのアクセスを制御するためのＵＲＬリストを拡張する方法について述べる。

2 フィルタリングソフトの現状

インターネットへのアクセスを制限するためのフィルタリングソフトは、ページ製作者等がページに書き込んだジャンルや表現の程度に関する識別情報のラベルとページ閲覧者等が設定した情報とを比較してアクセスの制限をしたり、フィルタリングソフト会社が作成したＵＲＬのリストを用いてアクセスの制限をしたりする。また、単語やフレーズの辞書を持ち、辞書内の語とページ内に出現する語を比較することにより、制限するか否かの判断をするものもある。ページ内のラベルを利用する方法は、適切にラベルが貼られていれば確かであるが、現状では、適切なラベルが貼られているページはまだ少ない。また、辞書を使う方法は、適切な辞書を作成することが困難であり、また、単語やフレーズによる機械的な判断は間違えることも多い。また、画像のみのページでテキストのないページも存在し、そのようなページには対処できない。このため、フィルタリングソフトは、ＵＲＬリストを必要とするが、新サイトの出現や既存サイトの移転などが頻繁におこるため、そのメンテナンスには多大な労力が必要とされている。

3 ページ内リンクリスト情報を利用したURLリスト拡張手法

インターネット上には、様々なリンク集的なページが存在し、ページ製作者によってメンテナンスが行われている。これを利用して、ＵＲＬリストを拡張する方法を提案する。

3.1 手法の概要

次のような手順で、ＵＲＬリストに追加するＵＲＬを決定する。

(1) ページ内リンクリストの作成
まず、ロボットで収集したWebページのＨＴＭＬファイルから、ＨＴＭＬタグを利用して、リンク先のＵＲＬの一覧を作成する。これをページ内リンクリストと呼ぶことにする。

(2) リンク集ページの推定
次に、規制対象ジャンルのリンク集を見つける必要がある。ここでは、あらかじめ規制対象ジャンルのＵＲＬリストをある程度人手で作成しておき(これを初期ＵＲＬリストと呼ぶ)、ページ内リンクリストと比較し、初期ＵＲＬリスト中のＵＲＬをある数以上含むページをリンク集と見なすことにする。

(3) 各リンク先ＵＲＬの評価値の算出
リンク集と判断したページのページ内リンクリス

URL database expansion method for web filtering by using the link lists in web pages
Toshiro MAKINO, Masayuki SUGIZAKI, and Hirohito INAGAKI
NTT Cyber Solutions Laboratories, NTT Corporation

ト中のＵＲＬのうち、初期ＵＲＬリストに存在しないＵＲＬに関して、評価値をつける。

(4) 追加するＵＲＬの決定
ロボットで収集したすべてのページで上記の（１）～（３）を行い、評価値がある閾値以上のＵＲＬを追加するＵＲＬとする。

3.2 評価値の算出方法

評価値の算出に関しては、以下の点を考慮した。

- １ページ内に複数のジャンルのリンクを掲載しているリンク集が存在すること。
- リンク集ページ内にも、リンク集とは無関係なリンクが存在すること。

１点目に関しては、まず、ページ内リンクリスト中のＵＲＬのうち、初期ＵＲＬリストに存在する存在するＵＲＬと存在しないＵＲＬの割合を考え、存在するＵＲＬの割合が高い場合に、存在しないＵＲＬの評価値が高くなるようにした。これにより、複数ジャンルのリンク集が存在するページでは各ＵＲＬの評価値が小さくなることが期待できる。また、複数ジャンルのリンク集が存在する場合、目的のジャンル以外のジャンルはページによって異なることが期待できるので、ページ毎の評価値の和を最終的な評価値とすることにより、さらに他ジャンルのＵＲＬの評価値を相対的に小さくすることができる。これは２点目の無関係なリンク先の評価値を下げるためにも有効である。２点目に関しては、さらに各サイトの被リンク数をカウントし、リンク集からのリンク数との割合を考え、目的のジャンルのリンク集以外からの被リンク数の割合が高いサイトに関しては、評価値が低くなるようにした。上記の考え方に従い、以下の式により評価値 v を求めた。

$$v = \frac{\sum_{i=1}^{M} \frac{b_i}{n_i}}{N}$$

ただし、b_i は、あるページ内リンクリスト i 中の初期ＵＲＬリストに含まれるＵＲＬの数、n_i は、ページ内リンクリスト i 中のＵＲＬの数、M はそのＵＲＬを含むリンク集ページの数、N は、そのＵＲＬの被リンク数である。

4 実験と考察

ロボットで収集したWebページからランダムサンプリングした７０万強のページを対象に、フィルタリングソフトで必ず規制対象とされるアダルトジャンルに関して、本手法によりＵＲＬリストの拡張の実験を行った。初期ＵＲＬリスト中のＵＲＬの数は１９７、初期ＵＲＬリスト中のＵＲＬを２つ以上含むページをリンク集ページと見なした。結果は、リンク集として判定されたページ数が３２５、閾値を０とした場合の追加ＵＲＬ候補数は３７７９であった。図１に閾値を変化させた場合の追加ＵＲＬ候補の数および、そのうち実際にアダルトジャンルのＵＲＬであった数、接続ができなかったＵＲＬの数を示す。なお、評価値の絶対値は初期ＵＲＬリストによって大きく変化するので、図では最大値を１として割合で表示してある。

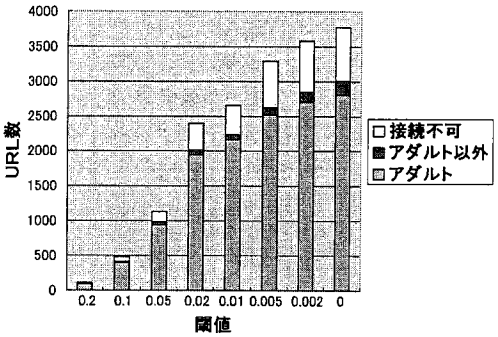

図１ 閾値による URL 数の変化

閾値に関わらず接続不可のＵＲＬを除いて求めた適合率は 0.93 以上あり、本手法で有効にＵＲＬの拡張ができることが確認できた。また、閾値が 0.2,0.1 の場合に、一時的に適合率が低下するが、それ以降は閾値を下げるとともに適合率が低下しており、本評価値の設定方法が、無関係のＵＲＬを取り除く効果を持つことが確認できた。一時的な適合率の低下は、同一作者が作成したと思われるほぼ同じ内容のアダルトリンク集ページからリンクされているアダルト以外のページで他のページからあまりリンクされていないページが存在し、そのページの評価値が高くなってしまっているのが原因であった。今後、処理ページ数を増やして、同様の問題が生じるか否かの検証を行うとともに、よりよい評価値の算出方法を検討したい。

5 おわりに

ロボットで収集したWebページ内のリンク情報を利用して、アクセス制御用のＵＲＬリストを拡張する方法を提案した。本手法は再帰的に適用できるため初期ＵＲＬリストが小さくても大量のＵＲＬリストを自動的に作成できる。