3U-4

# 有害サイトフィルタリングのための リンク構造に基づくブラックリストの拡張について

永田雄大　大囿忠親　伊藤孝行　新谷虎松

名古屋工業大学大学院 工学研究科情報工学専攻

e-mail: {yuta, ozono, itota, tora}@ics.nitech.ac.jp

## 1 はじめに

有害サイトフィルタリングでは，有害と判断されるWebページをリストアップし，アクセスできないようにするブラックリスト手法が中心となっている [1]．しかし，日々増加し続けるWebに対して，ブラックリストの維持・拡張は困難な作業と言える．本稿では，Web構造マイニング技術を応用することで，効率的にブラックリストを拡張する手法を提案する．有害サイトは相互リンクなどの密なリンク構造を持つ事に着目し，同一トピックを抽出するWebコミュニティ発見アルゴリズムを適用し，新たなブラックリストの候補を発見する．ブラックリストの候補の中には，有害サイトと無害なサイトとを結ぶブリッジの役割を果たすページが存在する．そこで，ブラックリストの候補からブリッジを除去することで，確認する手間を省くことができる．テキスト情報に頼ることがないので，画像のみを多用するサイトの発見，さらには，新語や流行語などのWebの流行の変化への対応が期待できる．

以下に，本稿の構成を示す．まず第２章で，有害サイト特有のリンク構造と基本的アイデアについて述べ，ブラックリスト拡張手法について説明する．次に第３章で，提案手法について評価実験を行い，第４章にて，本手法を考察する．最後に第５章で，本稿をまとめる．

## 2 有害サイトのWebコミュニティ

### 2.1 リンク構造の特徴

本稿では，リンク解析のみに頼ったアルゴリズムを採用する．なぜならば，有害サイトの場合，画像を多用しテキスト情報が乏しいことがあり，テキストに基づくコンテンツ解析は期待できない．従って，コンテンツは考慮せずリンク解析のみで有害サイトを発見する．

有害サイトのリンク構造に多く見られる特徴を以下にまとめる．(1) 有害サイト同士は，互いに密なリンク構造を持つ [1]. (2) 有害サイトと無害なサイトを結ぶブリッジの役割を果たすWebページが存在する．上記の考察をもとに，ブラックリストの拡張手法の基本的なアイデアを次に説明する．

有害サイト同士は密なリンク構造を持つことから，HITSアルゴリズム [2] を適用することで有害サイトのWebコミュニティを抽出することが可能であると考える．HITSアルゴリズムとは，Web特有のハイパーリンク情報を解析し，特定のトピックに関するページ集合(authorityとhubページからなるWebコミュニティ)を抽出するアルゴリズムである．HITSアルゴリズムでは，密にリンクを張り合うページ間において，authorityとhubが高いスコアとなる．従って，既存のブラックリストのページとリンク関係にあるページ集合において，authorityとhubのスコアが高いページは，有害サイトのWebコミュニティでありブラックリストの候補として考えることができる．このブラックリスト候補には，有害サイトと無害なサイトを結ぶブリッジの役割を果たすWebページを含む可能性がある．そこで，ブリッジとしてあらかじめ予測できるWebページをストップURLとしてデータベース化し，ブラックリストの候補集合から除去する．ブリッジの役割を果たすサイトとして，Yahoo![1]などの著名なサイトが多く観察された．

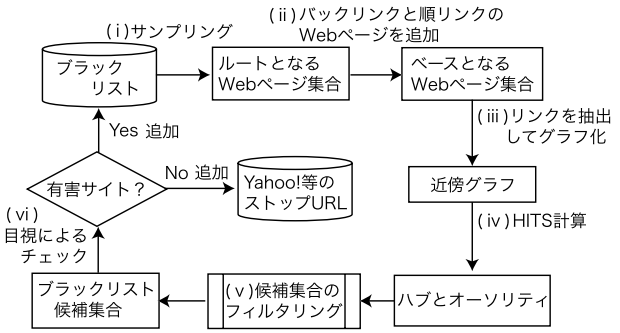


図 1: ブラックリスト拡張プロセスの流れ

### 2.2 ブラックリスト拡張プロセス

提案するブラックリスト拡張手法では，次の (i) から (vi) のプロセスを経てブラックリストを追加していく．図 1 にプロセスの流れを示す．

(i) ブラックリストの Web ページ集合 $B$ から，一定数 $t$ の Web ページをサンプリングし，ルート集合 $R_B$ を作成する．(ii) ルート集合 $R_B$ に含まれる Web ページからリンクされている Web ページ，およびルート集合 $R_B$ に含まれる Web ページにリンクしている Web ページを最大 $d$ 件収集した後，ルート集合に追加してサイズ $n$ のベース集合 $S_B$ を作成する．(iii) ベース集合 $S_B$ のページ間のリンク関係をすべて洗い出し，隣接行列を作成する．このとき，同一ドメイン間のリンク関係は除去する．(iv) ページ間のリンク関係を示す隣接行列をもとに HITS アルゴリズムを適用する．上位スコア一定数の authority と hub を抽出し，ブラックリストの候補集合とする．(v) ストップ URL，および既存のブラックリスト中の URL を候補集合から除去する．(vi) 候補集合のページを目視で確認し，有害サイトならばブラックリストに追加し，そうでないならストップ URL データベースに追加する．

ブラックリストの候補集合に残る無害なサイトは，今後もブラックリストの候補にあがるブリッジの可能性があるので，ストップ URL としてデータベース化する．ストップ URL を拡充することで，ブリッジとなる Web ページの除去精度の向上が期待できる．以上のアルゴリズムを，図 2 に示す．

## 3 評価実験

前章で述べた手法をもとにブラックリスト拡張システムを Java を用いて実装し動作させた．システムに入力

A method of Black List expansion based on Web structure for Web Filtering

Yuta NAGATA, Tadachika OZONO, Takayuki ITO, Toramatsu SHINTANI

Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology.

[1] http://www.yahoo.co.jp

**Candidacy Selection**
$B$ : ブラックリストのページ集合.
$Bridge$ : ストップ URL のページ集合.
$t, d, c, k$ : 自然数.
Iterate($G, k$) : グラフ G における
HITS 反復計算 k 回後の各ページの $x, y$.
$S_B$ := **Subgraph**($B, t, d$).
$G[S_B]$ : $S_B$ のページ間をリンクしたサブグラフ.
$G_B$ : $G[S_B]$ から同一ドメイン間リンクを消去したグラフ
$(x, y)$ := Iterate($G_B, k$).
$Authorities$ : $x$ の上位 $c$ 件のページ集合.
$Hubs$ : $y$ の上位 $c$ 件のページ集合.
$C_B := Authorities \cup Hubs$.
$C_B := C_B - Bridge$.
$C_B := C_B - (C_B \cap B)$.
**Return** $C_B$.

**Subgraph**($B, t, d$)
$R_B$ : $B$ から無作為に $t$ 件抽出したページ集合.
Set $S_B := R_B$.
**For** each page $p \in R_B$.
$\Gamma^+(p)$ : $p$ からリンクされるすべてのページ集合.
$\Gamma^-(p)$ : $p$ にリンクするすべてのページ集合.
$\Gamma^+(p)$ のすべてのページを $S_B$ に追加.
**If** $|\Gamma^-(p)| \preceq d$, then
$\Gamma^-(p)$ のすべてのページを $S_B$ に追加.
**Else**
$\Gamma^-(p)$ から無作為に $d$ 件を $S_B$ に追加.
**End**
**Return** $S_B$

図 2: 候補選択アルゴリズム

する URL は，コミュニティを作る基となるものであり，代表的な有害サイトであることが望ましい．本稿では，財団法人インターネット協会[2] から提供されているレイティングサービス SafetyOnline において，有害レベルの高いサイト 1,000 件をブラックリストとして扱った．ベース集合を作成する時のバックリンクを求める際には，Google[3]のリンク検索を用いた．$t = 100, d = 50$ とした場合のベース集合のサイズ $n$ は，概ね 2,000 から 3,000 の間であった．また，あらかじめストップ URL として著名なサイトのページをリストアップした．ストップ URL の代表例として，Yahoo!などのポータルサイト，Google などの検索サイト，および Windows Media Player[4] などのメディアプレイヤーのサイトがあげられる．

図 3 に実験の結果を示す．ある試行において得られた authority のスコア上位 500 件中の有害サイトの割合についてまとめ，有害サイトが Web コミュニティとして抽出できているか調査した．図 3 は，スコアの上位から 100 件毎に分割した際の，それぞれ 100 件中の有害サイトの割合について示している．また，ストップ URL によってブリッジを除去する前と後についてまとめ，ストップ URL の有効性について調査した．図 3 からわかるように，スコアが上位になるほど有害サイトの割合が高くなっている．これは，HITS アルゴリズムによって，ページ集合中の authority スコアが高いページが有害サイトの Web コミュニティとして抽出できていることを示している．また，ストップ URL を除去した場合に，有害サイトの割合が高くなっていることから，ストップ URL が有効に働いていることがわかる．

[2]http://www.iajapan.org/
[3]http://www.google.co.jp/
[4]http://www.microsoft.com/windows/windowsmedia/

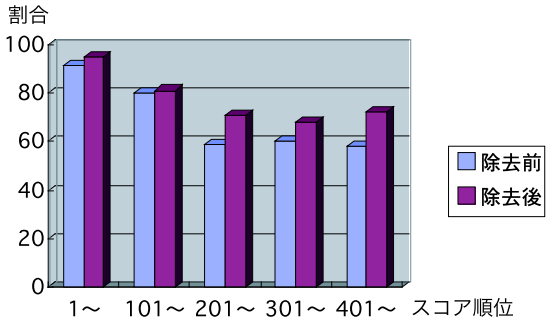


図 3: authority 上位 500 件中の有害サイトの割合

## 4 考察

前章の評価実験をもとに本手法の考察をまとめる．

評価よりブラックリスト候補の上位スコアにおいて，有害サイトの Web コミュニティが得られたと言える．下位スコアにおいては，既に Web 上にページが存在してない場合が多く，リンク切れによって得られた候補であった．有害サイト同士は互いに密なリンク関係を定期的に更新していることから，既に存在しないページはスコアが低くなり，リンク構造が保たれている既存のページ同士はスコアが高くなっていると考えられる．よって，本手法は無害なサイトだけでなく，リンク切れのページも除去できていると言える．

また，ブリッジの存在に着目し，著名なサイトをストップ URL として採用した．その結果，候補集合から無駄なページを除去するフィルタリングとしての効果を得る事ができた．さらに，ブラックリストの候補にあがるブリッジは，他の試行においても候補として出現していたことから，ストップ URL をさらに拡充する事でフィルタリング精度の向上が期待できる．

## 5 おわりに

本稿では，Web 構造マイニングの研究をもとにブラックリストを効率よく拡張するための手法について述べた．本手法の特徴は，以下の二点である．(1) 有害サイト同士の密なリンク関係をもとに，Web コミュニティ発見アルゴリズム HITS を適用した．(2) ブリッジの役割を果たすサイトをストップ URL として除去し，ブラックリスト候補における無駄なページの効果的な削減を行った．それぞれの特徴について評価実験に行い，本手法の有効性を確認した．本手法は，コンテンツ情報に頼らずリンク構造のみを用いて Web コミュニティを発見していることから，新語や流行語の影響が少なく Web の変化に対応可能である点，テキスト情報が乏しいページを発見することが可能である点，などが利点としてあげられる．また本稿は，Web コミュニティの発見という Web 構造の解明を行っていると言える．有害サイトのリンク構造がより明らかになれば，ブラックリスト拡張のための Web クローラーを作成することも期待できる

## 参考文献

[1] Chen Ding, Chi-Hung Chi, Jing Deng, Chun-Lei Dong, “Centralized content-based Web filtering and blocking: how far can it go?”, Proc. of IEEE SMC ’99, Vol.2, pp.115-119,1999.

[2] Jon M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, Journal of the ACM, Vol.46, No.5, pp.604-632, 1999.