D-020

WWW 画像検索システムにおける有害画像フィルタリング手法

A Method of Filtering Hazardous Images on WWW Image Search Systems

小泉 大地† 獅々堀 正幹† 中川 嘉之‡ 柘植 覚† 北 研二†
Daichi Koizumi Masami Shishibori Yoshiyuki Nakagawa Satoru Tsuge Kenji Kita

1. はじめに

近年，インターネットの普及に伴い，急速に Web サイト数が増大しているが，Web サイトの中には未成年者にとって不適切な情報が多く存在している．この問題に対処するため，利用者がアクセスできる情報を制限するフィルタリングシステムが開発されている[1][2]．特に，WWW 画像検索システムは，教育現場において資料収集のために頻繁に用いられているにもかかわらず，一般的なキーに対する検索結果内にも多くの有害な画像が表示されてしまうため，フィルタリング処理の適用が望まれている．現在，既存の WWW 画像検索システムでは，有害画像の URL をデータベース化することでフィルタリングするものも存在する．しかし，有効な URL がデータベース化されていないため，高精度なフィルタリングは実現できていない．

そこで本稿では，URL をパス毎の出現頻度を用いて重みづけを行うことにより，有害性の高い URL を識別しフィルタリングする手法を提案する．フィルタリングに必要な URL データベースは，数十個のキーワード群を用意しておくだけで自動構築できる．

2. URL の部分マッチングによるフィルタリング手法

2.1 本手法を用いたフィルタリングシステム

本フィルタリング手法を用いることにより，図 1 に示すようなフィルタリングシステムを構築できる．このフィルタリングシステムは URL データベースの構築処理とフィルタリング処理に分けられる．以下に 2 つの処理についてそれぞれ説明する．

URL データベース構築処理：

既存の WWW 画像検索システムを用いて，有害画像の URL データベース（Hazardous URL DB）を構築する．まず，WWW 画像検索システムに有害な画像を象徴するキーワード（Hazardous キーワード）群を入力して，検索された URL を Hazardous URL DB に登録する．その後，このデータベースの各 URL に出現頻度を付与する．

フィルタリング処理：

ユーザが検索質問を入力すると，WWW 画像検索システムが出力する検索結果内の URL と Hazardous URL DB 中の URL との部分マッチングを行い，Hazardous と判定した URL にリンクする画像にアクセスできないようにする．

2.2 URL 出現頻度の抽出

URL に着目して有害画像のフィルタリングをするため

† 徳島大学
‡ インフォコム株式会社

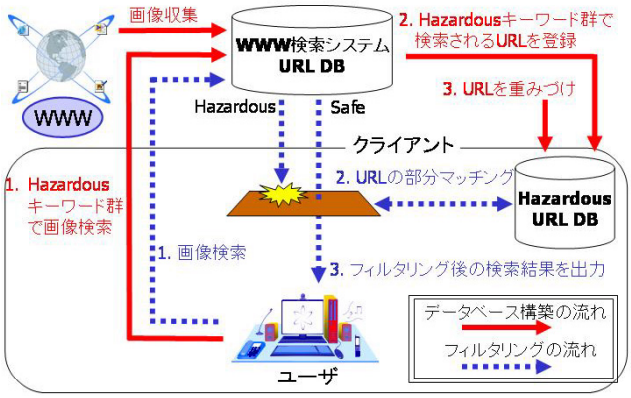

図 1. 本手法を用いたフィルタリングシステム

には，WWW 空間全体を基準に正規化した URL 出現頻度が必要になる．本手法では，Hazardous URL DB 中の部分 URL に対する WWW 空間全体における URL の出現頻度として，既存の WWW 検索エンジンの URL 検索機能を用いて正規化を行った．

手順 1： Hazardous URL DB 中の URL をパス毎に区切り，部分 URL の出現頻度を求める．

手順 2：手順 1 で求めた各部分 URL に対して，WWW 空間全体での出現頻度（大域的出現頻度）を求める．この値は，既存の WWW 検索システムの URL 検索機能を用いることで求める．

手順 3：手順 2 で求めた大域的出現頻度を用いて，各部分 URL を正規化した出現頻度（部分 URL の有害度 H_{url}）を求める．計算方法は式(1)で求める．

$$H_{url} = \frac{\text{Hazardous URL DB 内の部分URL出現頻度}}{\text{部分URLの大域的出現頻度}} \quad (1)$$

2.3 URL の部分マッチング

部分 URL の有害度 H_{url} を用いて，検索結果中の有害画像をフィルタリングする手順を示す．

手順 1： H_{url} の閾値 T を設定する． T を変動させることにより，フィルタリングのレベルを設定することが可能である．

手順 2：検索結果中の画像にリンクする URL と，Hazardous URL DB の URL を始端部から順に部分マッチングを行う．

手順 3： マッチした部分 URL 毎に H_{url} の判定を行う． T 以上となる部分 URL が Hazardous URL DB に登録されていれば，その URL を有害（Hazardous）とみなし，登録されていなければその URL を無害 （Safe）とみなす．

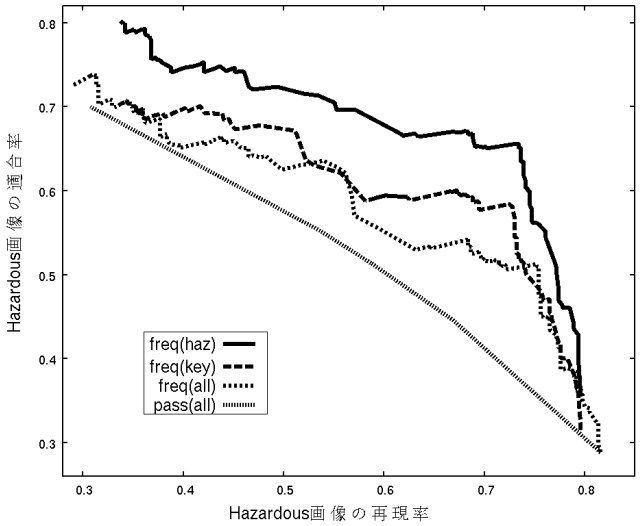

図 2. Hazardous 画像の再現率・適合率

3. Hazardous キーワードの選定

本フィルタリングシステムでは，Hazardous キーワードで検索された URL は，すべて Hazardous な情報を含む URL と仮定して Hazardous URL DB に登録されてしまう．しかし，この URL の中には Safe なものも存在するため，部分マッチングで Safe な URL を誤って Hazardous と識別してしまう可能性がある．そこで，フィルタリングに有効な Hazardous キーワードだけを自動で選定する手法を考案し，改良を加えた．

本選定手法は，Hazardous キーワード毎に検索結果内の画像にリンクする HTML 内のコンテンツを解析することにより，有害性の低いキーワードを除外する．以下に選定手法の手順を示す．

手順 1：Hazardous キーワード毎に検索結果内の画像にリンクするページを取得する．

手順 2：ページ内に出現する単語の中から，検索キーワード以外の Hazardous キーワードの異なり数の平均値を求める．

手順 3：求めた平均値を Hazardous キーワードの有害度と呼び，有害度の高い上位のキーワードのみを Hazardous キーワードとして選定する．

4. 評価実験

4.1 実験条件

Google Image Search を用いて 3 種類の Hazardous URL DB と評価用 DB を作成した．まず，54 個の Hazardous キーワードで検索し，検索結果上位 100 件の URL 計 4,189 件を DB *all* に登録した．次に，DB *all* 中から Hazardous なものだけを人手で取り出し，2,396 件の URL を DB *haz* に登録した．また，Hazardous キーワード選定手法を用いて，54 個の Hazardous キーワードの中から選定した上位 40 件のキーワードを用いて検索した 3,061 件の URL を DB *key* に登録した．最後に，Hazardous キーワードとは別に有害な画像が検索される可能性がある“看護婦”や“制服”と

表 1.手法における F 尺度の平均値

	freq(*haz*)	**freq(*key*)**	**freq(*all*)**	**pass(*all*)**
haz	0.5992	0.5551	0.5498	0.5032
saf	0.8648	0.8469	0.8295	0.8243

いった 27 個の評価用キーワードで検索を行い，検索された URL 計 2,639 件を評価用 DB とした．更に，評価用 DB 中の URL を人手で Hazardous と Safe に分類し，456 件の Hazardous URL と 2,183 件の Safe URL を得た．

4.2 Hazardous 画像の再現率・適合率

評価用 DB に対して DB *all*，DB *haz*，DB *key* を用いてフィルタリングを行い， Hazardous 画像の再現率 R_{haz}，適合率 P_{haz} を求めた． R_{haz} は評価用データ中の全 Hazardous 画像を正しくブロックできた割合を表し， P_{haz} はブロックした画像の中で本当に Hazardous 画像であった割合を表す．閾値 T を 0.0～1.0 まで 0.0001 毎に変化させた結果，図 2 に示す再現率・適合率曲線を得ることができた．図中の“freq”は本フィルタリング手法を用いた結果であり，“pass”は URL の出現頻度を考慮しない部分マッチングを用いた結果である．また，再現率と適合率を総合的な観点から 1 つの値により評価するために F 尺度を求めた．図 2 のグラフに対し，再現率を 0.05 毎に区切った計 101 点の F 尺度を計算し，その平均値を求めた．各手法の F 尺度の平均値を表 1 に示す．

実験結果より，“pass(*all*)”に比べ“freq(*all*)”が高い値を示していることから，正規化した頻度を用いたフィルタリング手法が有効であるといえる．また，“freq(*all*)”に比べ“freq(*key*)”の値が向上しているので，キーワード選定を行った結果，Safe URL が少ない URL データベースの構築に成功しているといえる．最終的に，人手でレイティングした結果である“freq(*haz*)”に最も近い精度が“freq(*key*)”であり，本手法の有効性が確認できる．

5. まとめ

本稿では，URL をパス毎に重みづけを行うことにより，有害性の高い URL を部分的に識別しフィルタリングする手法を提案した．また，URL データベース構築に必要な Hazardous キーワードの選定手法を考案した．今後は，関連キーワード自動収集手法[3]を用いた Hazardous キーワード拡張によるフィルタリング精度の向上を行いたい．

謝辞：本研究の一部は，科学研究費補助金基盤研究(B)(17300036), 科学研究費補助金基盤研究(C)(17500644)を受けて行われた．

参考文献

[1] 井ノ上, 帆足, 橋本: 文書自動分類手法を用いた有害情報フィルタリングソフトの開発, 信学論 D-Ⅱ, J84-D-Ⅱ, No.6, pp.1158-1166, 2001.

[2] 武者, 広池, 森本, 松田: WWW 有害情報のフィルタリングのための画像判別手法, FIT 2002, I-82, pp163-164, 2002.

[3] 竹安, 獅々堀, 柘植, 北: 出現 URL の類似性に着目した WWW 空間からの関連キーワード自動収集手法, 言語処理学会第 11 回年次大会, P4-5, pp691-694, 2005.