# D-030　WWW 画像検索システムを用いた有害サイト URL データベースの構築手法

## A Method to Construct URL Databases of Hazardous Web Site by Using WWW Image Retrieval Systems

中川 嘉之 † 獅々堀 正幹 † 小泉 大地 † 柘植 覚 † 北 研二 ‡
Yoshiyuki Nakagawa Masami Shishibori Daichi Koizumi 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，有害な Web サイトの 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，既存の WWW 画像検索システムを用いてフィルタリングに有効な URL データベースを効率的に構築する手法を提案する．本手法を用いると，数十個のキーワード群を用意しておくだけで，各検索システムに適合した URL データベースを構築できる．

### 2. WWW 有害情報フィルタリングシステム

既存のフィルタリングシステムは大きく分類すると，(1) アクセス制限する URL 一覧をデータベース化する URL チェック方式，(2) 利用者がアクセスした Web サイトに含まれる単語や画像といったコンテンツを解析するコンテンツチェック方式に分けられる．

URL チェック方式では，事前に蓄積した URL を持つ Web サイトにしかアクセス制限ができないため，適切かつ大量の URL を事前に登録する必要があり，そのような大量の URL を如何に効率的に収集するかが問題となる．一方，コンテンツチェック方式では，事前に URL を登録しなくてもよいが，コンテンツの解析が Web サイトのアクセス時に行われるため，時間的コストが要求されるといった問題点がある．特に本稿で対象とする，WWW 画像検索システムの検索結果をフィルタリングする場合，検索結果には複数のサムネイル画像が含まれており，画像は文書に比べ解析時間が要求されるので高速な処理が可能な URL チェック方式が実用的である．ただし，WWW 画像検索システムの検索結果に含まれる有害画像を適切にフィルタリングできるように URL データベースを構築するためには，次のような問題が生じる．

**問題 1：** WWW 空間全体を対象に URL を収集する場合，自動収集ロボットを用いると大量の URL を収集するのに莫大な時間を要する．

**問題 2：** Web サイトが有害であるか否かの判定（レイティング）を行うために，各 Web サイトのコンテンツを人手で判定するには多大な労力がかかる．

**問題 3：** 各検索システムでは，個別に Web サイトの収集アルゴリズムを用いているため，検索システム毎に検索結果が異なっている．従って，WWW 空間全体から収集構築した URL データベースは汎用性に欠ける．

†徳島大学工学部
‡徳島大学高度情報化基盤センター

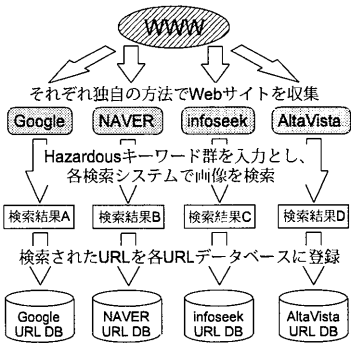


図 1: WWW 画像検索システムを用いた URL 収集

本稿では，これらの問題点に対応した URL データベース構築手法を提案する．

### 3. WWW 画像検索システムの利用

#### 3.1 WWW 画像検索システムを用いる利点

有害情報フィルタリングに用いる URL データベースを構築する際に，WWW 画像検索システムを用いることにより，2. で述べた問題点を以下のように解消できる．

**対処 1：** 有害な画像が掲載されている付近には，有害な画像を象徴するキーワード（以後，Hazardous キーワードと記す）が存在することが多い．そのため，WWW 画像検索システムにおいて，Hazardous キーワードを入力とすると有害画像が大量に検索され，その画像にリンクされている URL を収集することで，簡単に有害サイトの URL 候補を大量に収集できる．

**対処 2：** WWW 画像検索システムの検索結果には画像が表示されるので，その画像を一目見るだけでレイティングすることができる．

**対処 3：** WWW 画像検索システム毎に Hazardous キーワードを入力として収集した URL データベースは，各検索システムの収集アルゴリズムが反映されており，各検索システムに適合した URL データベースが構築できる．URL データベースの更新は，索引の更新があった検索システムに対してのみ行えばよい．

#### 3.2 URL データベースの構築

WWW 画像検索システムを用いた URL 収集方法を図 1 に示す．まず，WWW 画像検索システム毎に同一の Hazardous キーワード群で検索を行う．そして，各検索システムで検索された画像にリンクされている URL をデータベース化する．今回は個々の画像のレイティングを行わず，Hazardous キーワードで検索された画像はすべて有害と仮定してすべての URL をデータベースに登録した．

表 1: URL データベースに登録された URL 数

| ISE1 | ISE2 | ISE3 | ISE4 | ISE5 |
|---|---|---|---|---|
| 17,198 | 41,710 | 19,617 | 3,643 | 6,337 |

表 2: 評価用データの詳細

| | ISE1 | ISE2 | ISE3 | ISE4 | ISE5 |
|---|---|---|---|---|---|
| Hazardous | 550 | 696 | 643 | 70 | 56 |
| Safe | 2,150 | 2,025 | 1,978 | 2,097 | 2,229 |

URL のマッチングには，登録された URL の一部を比較する方法を用いた．まず URL を，(1)“http://aaa/bbb.html”，(2)“http://ccc/ddd/……”の 2 通りに分類した．(1) の形式の URL からは“aaa”を，(2) の形式の URL からは“ccc/ddd”を抽出し，抽出部分と一致していればアクセスを制限した．有害な情報が含まれる Web サイトはそのサイトのトップ以下全体が有害であると考えられる．これにより，データベースに登録された URL 数より多くの Web サイトを制限できる．

## 4. 評価実験

### 4.1 実験条件

まず既存の WWW 画像検索システムとして，Google, NAVER, infoseek, AltaVista, Yahoo を用いて 54 個の Hazardous キーワードで検索し，各 URL データベースを構築した．各 URL データベースに登録されてある URL 数を表 1 に示す．表中の ISE1〜ISE5 はそれぞれ，Google, NAVER, infoseek, AltaVista, Yahoo を表している．

次にこの各検索システムにおいて，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”といった 27 個の評価用キーワードで検索を行い，各検索システム毎に検索結果上位 100 件の画像とその URL を評価用データとした．更に，各評価用データを人手で判定し，性的描写がある場合は Hazardous，性的描写がない場合は Safe の 2 種類に分類した．各検索システムの評価用データ中の Hazardous 画像数及び Safe 画像数を表 2 に示す§．AltaVista 及び Yahoo の Hazardous 画像数が他の検索システムに比べて少ないのは，AltaVista, Yahoo の検索システム自体で有害情報のフィルタリングを行っており，評価にフィルタリング後の検索結果を用いたためである．

### 4.2 Hazardous 画像の再現率

5 つの評価用データに対して，各々に 5 つの URL データベースを用いてフィルタリングを行い，式 1 に示す Hazardous 画像の再現率（$Recall_{haz}$）を求めた．$Recall_{haz}$ は，検索した Hazardous 画像を正しくフィルタリングした割合を表す．

$$Recall_{haz} = \frac{\text{正しく } Hazardous \text{ と分類された画像数}}{\text{全 } Hazardous \text{ 画像数}} \quad (1)$$

Hazardous 画像の再現率を表 3 に示す．表中の DB は URL データベース，DATA は評価用データを表す．結果から，Google, NAVER, infoseek では，各評価用データに同じ検索システムの URL データベースを用いた方が優れたフィルタリング性能を発揮していることが分かる．従って，WWW 画像検索システムを用いて構築した各 URL データベースは，同一システムの検索結果に対する有害情報のフィルタリングに有効であると言える．AltaVista 及び Yahoo で再現率が低いのは，フィルタリング後の検索結果を用いており，URL データベースに登録された有害 Web サイトの URL 数が少ないためだと考えられる．

§Hazardous 画像数と Safe 画像数の和が各検索システム毎に異なっているのは，キーワードによっては検索結果が 100 件に満たなかったためである．

表 3: Hazardous 画像の再現率

| DB ＼ DATA | ISE1 | ISE2 | ISE3 | ISE4 | ISE5 |
|---|---|---|---|---|---|
| ISE1 | 0.84 | 0.37 | 0.13 | 0.03 | 0.11 |
| ISE2 | 0.35 | 0.92 | 0.19 | 0.04 | 0.14 |
| ISE3 | 0.18 | 0.07 | 0.80 | 0.01 | 0 |
| ISE4 | 0 | 0 | 0 | 0.33 | 0.29 |
| ISE5 | 0.01 | 0.01 | 0 | 0.36 | 0.30 |

表 4: Safe 画像の再現率

| ISE1 | ISE2 | ISE3 | ISE4 | ISE5 |
|---|---|---|---|---|
| 0.77 | 0.70 | 0.72 | 0.92 | 0.90 |

また，フィルタリングに失敗した原因を考察するため，Hazardous 画像が存在する Web サイトのコンテンツを調査した．その結果，次の 2 つの原因に分類できた．(1)Hazardous キーワードが用いられていない．(2)Hazadous キーワードが存在するが URL データベースに登録されていない．(1) の対策としては，上位サイトの URL とのマッチングを行う方法が考えられる．

### 4.3 Safe 画像の再現率

評価用データと同一の検索システムから構築した URL データベースの組み合せで，評価用データに対してフィルタリングを行い，式 2 に示す Safe 画像の再現率（$Recall_{saf}$）を求めた．$Recall_{saf}$ は，検索した Safe 画像に対してアクセスを許す割合を表している．

$$Recall_{saf} = \frac{\text{正しく } Safe \text{ と分類された画像数}}{\text{全 } Safe \text{ 画像数}} \quad (2)$$

Safe 画像の再現率を表 4 に示す．結果より，Google, NAVER, infoseek では 7 割程度，AltaVista, Yahoo では 9 割程度 Safe 画像に対してアクセスを許していることが分かる．誤って制限された Safe 画像が存在するサイトの多くは，コンテンツも安全な内容であった．これは，今回用いた URL マッチングの方法に問題があると考えられる．この改良策として，URL の一部を抽出するのではなく，“/”間の URL 毎に重みづけを行う方法が考えられる．

## 5. まとめ

本稿では，既存の WWW 画像検索システムを用いてフィルタリングに有効な URL データベースを効率的に構築する手法を提案した．評価実験では，WWW 画像検索システムを用いて構築した各 URL データベースは，同一システムの検索結果に対する有害情報のフィルタリングに有効であることを証明できた．今後の課題としては，URL マッチングの改良と画像のレイティングによる再現率の向上があげられる．

## 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.