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

小泉大地[†] 獅々堀正幹[††] 中川嘉之[†,☆]
柘植覚[††] 北研二[†††]

WWW 画像検索システムは検索キーに対する画像を WWW 空間から手軽に取得できるため，教育現場における資料収集ツールとして頻繁に利用されているが，一般的なキーに対する検索結果内に有害な画像が含まれるといった問題点がある．この問題点に対して，有害画像を含むページの URL をデータベース化し，検索結果の各 URL をチェックすることで有害画像をフィルタリングするシステムが存在する．しかし，フィルタリング精度を高めるために，大規模な URL データベースの構築，更新作業に多大な労力が必要になる．そこで本論文では，ユーザ・サイドで構築した URL データベースを用いて，既存の WWW 画像検索システムの検索結果から有害画像をフィルタリングする手法を提案する．本手法は有害画像を象徴する単語群をユーザ・サイドで用意するだけで検索サーバごとに URL データベースを自動構築する．そして，自動構築した URL データベースには URL のパスごとに有害度の重みづけを行い，有害性の高い URL を部分的に識別することで高精度なフィルタリングを実現する．さらに，URL データベース自動構築の際に用意する各単語の意味的多義性に着目し，ノイズ混入の可能性がある単語をあらかじめ自動選別することでフィルタリング精度を高める．既存システムとして Google Image Search を用い，有害な画像が検索される可能性のある 27 個の検索キーに対する評価実験を行った結果，F 尺度として約 70%のフィルタリング精度が得られた．

A Method of Filtering Hazardous Images on WWW Image Search Systems

Daichi Koizumi,[†] Masami Shishibori,[††] Yoshiyuki Nakagawa,[†,☆]
Satoru Tsuge[††] and Kenji Kita[†††]

WWW image retrieval systems are extremely useful for collecting educational materials from the WWW space. The search results, however, often include sexually explicit or pornographic contents, which are not suitable for educational purposes. Some filtering systems use simple URL databases to filter out hazardous images, but these are not always effective, since it is very hard to maintain constantly changing URL databases. In this paper, we propose a new filtering method using partial URL-based weighing scheme. The method first gathers a lot of hazardous URLs from search results of conventional WWW image retrieval systems by hazardous keywords. The method next decomposes a full URL into several partial paths, then estimates hazardous score for each partial path based on frequencies of paths. And it filters out hazardous images by distinguishing partial URL that there is possibility of hazardous. In addition, we focus on the ambiguity of hazardous keywords, and the method to select only suitable hazardous keywords is also proposed. Experimental results show that the proposed method can improve the accuracy of filtering out hazardous images from search results of conventional systems.

1. はじめに

インターネットは必要な情報を手軽に入手することが可能であり，年齢層を問わず様々な人に利用されるようになってきた．しかし，Web サイトの中には未成年者に対し有害な情報も多く存在しているため[1)]，利用者が閲覧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[2),3)]．特に WWW 画像検索システムでは検索キーに対する画像を WWW 空間から手軽に取得できるため，教育現場における資料収集

† 徳島大学大学院工学研究科
Graduate School of Engineering, Tokushima University
†† 徳島大学工学部
Faculty of Engineering, Tokushima University
††† 徳島大学高度情報化基盤センター
Center for Advanced Information Technology, Tokushima University
☆ 現在，インフォコム株式会社
Presently with Infocom Corporation

ツールとして頻繁に利用されているが，一般的なキーに対する検索結果内に有害な画像が含まれるといった問題点がある．

WWW 画像検索システムは画像のファイル名や画像の周辺に存在する単語，および，画像のキャプション情報などから，検索キーに関連する画像を検索することができる．代表的なものとして，Google[4]，goo[5]，AltaVista[6]，Yahoo[7] があげられる．このシステムで問題となるのは検索結果に有害な画像が表示されてしまうことである．Web 検索の場合には有害サイトが検索されても検索結果中にはそのサイトへのリンク，および検索キーが存在する近辺のテキストしか表示されない．そのため，有害な検索キーが存在する近辺のテキストが表示されたとしても有害性は低い．一方，画像検索では検索結果中にサムネイル画像として有害画像が表示されてしまうため，有害性が非常に高い．そこで，既存の WWW 画像検索システムでは有害画像のフィルタリング機能を提供している．

goo や AltaVista では検索キーに対して制限するフィルタリング機能を提供している☆．両検索システムでは有害画像が検索されるであろう検索キーに対して制限することで，有害画像を閲覧できないようにしている．しかし，検索キーによっては多くの無害な画像も制限してしまう点，制限されていない検索キーに対してはフィルタリングされない点などが問題点としてあげられる．

Google や Yahoo ではアクセス制限する URL 一覧をデータベース化し，データベースに登録された URL とマッチする画像に対して制限を行っている[8]．この方法では誤って無害な画像を制限してしまうことはないが，URL データベースの構築を人手で行わなければならないため，多大な労力が必要となる．そのため Google や Yahoo では，すべての有害画像の URL をデータベース化できておらず，多くの有害画像が表示されてしまっている．したがって，URL データベースを用いた検索システムでは，ユーザ・サイドにおいてさらなるフィルタリング処理の適用が必要である．

ユーザ・サイドでフィルタリング処理を施す研究としては，Web ページのコンテンツを解析することで有害情報を制限する手法[9),10)] がある．この手法は，あらかじめ設定した有害単語が Web ページのコンテンツ内に含まれている場合，アクセスできないようにする方式である．しかしコンテンツの解析は，Web ページにアクセスしたときに行われるため時間的コストがかかるといった問題点がある．

そこで本論文では，ユーザ・サイドで自動構築した URL データベースを用いて，既存の WWW 画像検索システムの検索結果から瞬時に有害画像をフィルタリングする手法を提案する．ユーザ・サイドで URL データベースを構築することにより，検索サイトごとにカスタマイズしたデータベースを構築でき，またフィルタリングサービスを提供していない検索サイトに対しても本手法を適用できる．本手法は有害画像を象徴する単語群をユーザ・サイドで用意するだけで検索サーバごとに URL データベースを自動構築する．そして，自動構築した URL データベースには，URL のパスごとに有害度の重みづけを行い，有害性の高い URL を部分的に識別することで高精度なフィルタリングを可能にする．データベース内の情報に対し識別を行うため時間的コストもかからない．さらに，URL データベース自動構築の際に用意する各単語の意味的多義性に着目し，ノイズ混入の可能性がある単語をあらかじめ自動選別することで，フィルタリング精度を高める．

以下，2 章で既存のフィルタリングシステムの特徴と問題点について述べ，3 章でブラックリスト方式をベースとする本フィルタリング手法を提案し，4 章で URL データベースを構築する際に準備する単語の選定手法を提案する．5 章において本手法の有効性を検証するため，Google Image Search を用いた評価実験を示し，その考察を述べる．最後に 6 章において，まとめおよび今後の研究課題について述べる．

2. WWW 有害画像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）ブラックリスト方式，（2）コンテンツチェック方式，（3）レイティング/ラベリング方式，（4）キーワードフィルタリング方式に分けられる．以下に各フィルタリングシステムの特徴と問題点を述べる．

2.1 ブラックリスト方式

ブラックリスト方式は有害画像の URL をデータベース化し，その URL にアクセスできないようにする手法である．登録された URL は確実にフィルタリングすることができ，無害な画像や有益な画像を遮断してしまう可能性が少ないという利点がある．そのため，有効な URL データベースが構築できれば，最も高精度なフィルタリングが実現可能である．しかし，検索システムの URL データベースは大規模であり，莫大な数の画像から有害画像を人手で判定するには多大な時間，労力がかかってしまう[11)]．また，検索システム

☆ 実際には，検索キーによる制限だけではなく，人手でレイティングを行い制限していると推測できるが，大きな特徴として検索キーによる制限があげられる．

の URL データベースは定期的に更新されるため，有害画像の URL データベースもそのつど更新しなければならならない．

Google や Yahoo が提供しているフィルタリング機能はブラックリスト方式が用いられている．両検索システムでは有害サイトの URL データベースが不十分なため精度が悪く，多くの有害画像が閲覧できてしまうのが現状である．したがって，大量の有害画像の URL をいかに効率的に収集するかがブラックリスト方式の課題である．

2.2 コンテンツチェック方式

コンテンツチェック方式はユーザがアクセスしたページに含まれる画像，Web ページのコンテンツを解析することによって，有害画像の自動判定を行う手法である[9),10)]．コンテンツ解析技術を用いて判定を行うため，ブラックリスト方式のように事前に人手で有害画像を見つけ出し登録を行う作業が不要となる．しかし，色情報や形状情報といった画像のコンテンツを表す特徴量を用いるため，抽象的なデータとなり高精度な処理は望めない．たとえば，肌色領域が多い画像を有害画像と判定してしまうと，多くの顔画像もフィルタリングされてしまい，現在の画像処理技術では有害画像の有無を判定するのは困難である．また，コンテンツの解析が Web ページ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もある．特に，WWW 画像検索システムの検索結果には複数枚の画像が出力されるため，それらすべての画像を一度に解析しなければならない．リンク先の文書を解析する場合であっても，出力された画像の枚数分のページを解析しなければならないため，解析に要する時間的コストが大きくなってしまう．したがって，高精度な解析技術の適用と解析時間の短縮が課題である．

2.3 レイティング/ラベリング方式

レイティング/ラベリング方式は，PICS[12)] のレイティング基準[13)] に基づいて構成されており，あらかじめ画像にコンテンツを表すラベルを付与しておくことで，情報受信者がそのラベルを参照してアクセス制限する手法である．各画像に対しラベリングが施されているため，コンテンツチェック方式に比べ無害な情報を制限する恐れがなく，ブラックリスト方式に比べて受信者が情報を取捨する際の選択肢が広いことなどが利点としてあげられる．しかし，情報発信者全員にレイティング作業を義務づけるのは困難であり，ラベリングされたページが WWW 上に少なければ，高精度なフィルタリングを実現することはできない．実際に，有害な画像を象徴するキーワードで検索した約 10,000 件の HTML 文書のラベルタグを調べたが，PICS に準じてセルフレイティングされた HTML 文書はわずか 200 件程度であった．このようなことから，正しくラベリングされていれば正確にコンテンツを解析することができるため，他の方式と組み合わせて用いるのが適当である．

2.4 キーワードフィルタリング方式

キーワードフィルタリング方式は WWW 検索システムで有害なキーワードを用いて検索を行った場合に，その検索キーを無効にする手法である．有害な検索キーに対してのみ制限を行うため，容易にフィルタリングシステムが構築できる．goo や AltaVista はキーワードフィルタリング方式を用いたフィルタリング機能を提供している．

キーワードフィルタリング方式を用いると，非常に多くの有害画像をフィルタリングすることが可能であるが，キーワードによっては多くの無害な画像も制限されてしまうのが欠点である．たとえば，「看護婦」のような有害な画像，無害な画像が同程度に検索されてしまうキーワードであっても，有害な検索キーとして登録されているため，すべての画像がフィルタリングされてしまう．逆に，「看護婦」と同程度の検索結果が表示される「女子高生」のようなキーワードでは，有害な検索キーとして登録されていないため，安全な画像とあわせて多くの有害な画像も表示されてしまう．したがって，キーワードフィルタリング方式を用いたフィルタリングシステムでは，高精度なフィルタリング処理を望むことはできない．

3. URL の部分マッチングによるフィルタリング手法

2 章より，WWW 画像検索システムの検索結果に対するフィルタリング処理には，ブラックリスト方式とコンテンツチェック方式が有効であるといえる．特にブラックリスト方式では有効なデータベースが構築できれば高精度なフィルタリングができ，高速な処理も可能であるため，フィルタリングに最も適していると考えられる．

そこで本論文ではブラックリスト方式をベースとするフィルタリング手法を提案する．本手法は数十個のキーワード群で自動構築できる URL データベースを用いて，未分類 URL のパスごとに有害度の判定を行うことによりフィルタリングする．これにより，有害画像の URL を完全に網羅しなくても，データベースに登録された URL 数以上のフィルタリング効果を発揮することができる．さらに，キーワード選定方式を

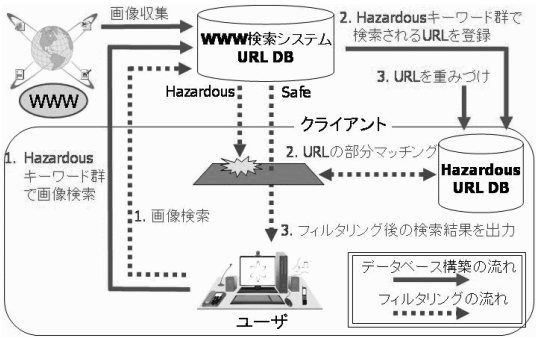

図 1 本手法を用いたフィルタリングシステム
Fig. 1 Outline of this filtering system.

用いて URL データベースの構築を支援する手法もあわせて適用する．この手法を用いれば，URL データベースの自動構築によるノイズ混入を緩和することが期待できる．

3.1 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．

3.1.1 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，検索された画像にリンクするページの URL を Hazardous URL DB に登録する．ここで，データベースに登録された URL に完全一致するものだけをフィルタリングした場合，適合性は高くなるが未登録の URL はフィルタリングできないため再現性が低くなってしまう．またサーバ全体を制限すると，再現性は高くなるが包括規制☆がかかってしまい適合性が低くなってしまう．実際には，URL は階層構造になっており，有害サイトが存在する場所も URL のパスごとに分岐していると考えられる．そこで，完全一致する URL だけやサーバ全体で制限を行うのではなく，各 URL のパスごとにフィルタリングを行う必要がある．なお，本論文では各 URL に対するパスごとの接頭文字列を部分 URL と呼び，部分 URL ごとにフィルタリングを行う．

しかし，各部分 URL 以下のコンテンツには，有害な情報，有害でない情報が混在している．そのため，部分マッチング（パスごとの文字列照合）で一致した URL をそのつど制限してしまうのでは，過剰なフィルタリングがおこってしまう．この過剰な規制を抑えるためには，各部分 URL 以下のコンテンツにどれだけ有害情報が含まれているかといった指標（以下，有害度と呼ぶ）が必要となる．本手法では部分 URL ごとに対応する有害度の値を付与する．なお，有害度の計算方法については 3.2 節で詳しく説明する．

3.1.2 フィルタリング処理

ユーザが入力した検索質問に対し，WWW 画像検索システムが出力する検索結果内の画像にリンクするページの URL と Hazardous URL DB 中の URL との部分マッチングを行う．部分マッチングの判定には，データベース構築の際に計算した有害度を用いる．これにより，未登録の URL に対しても制限することが可能となる．有害度が閾値以上となる部分 URL がデータベース内に登録されていれば，その URL にリンクする画像にアクセスできないようにする．このようにして，ユーザは有害画像がフィルタリングされた後の検索結果を閲覧することができる．

3.2 部分 URL ごとの有害度の計算

本手法では部分 URL ごとの有害度を用いて有害画像のフィルタリングを行う．以下に有害度の計算手順を示す．またこれらの手順に対応した例を**図 2** に示す．

手順 1：部分 URL の抽出
Hazardous URL DB 中のすべての URL に対し部分 URL を抽出する．

手順 2：部分 URL の出現頻度の抽出
Hazardous URL DB 内で各部分 URL が重複する頻度（部分 URL の出現頻度）を求める．URL は Hazardous キーワードを用いて収集されたことから，この出現頻度値が高い部分 URL ほど，有害情報を多く含んでいるといえる．

手順 3：部分 URL の大域的出現頻度の抽出
WWW 空間全体における各部分 URL 以下のコンテンツ数（部分 URL の大域的出現頻度）を求める．本来ならば実際に各部分 URL にアクセスし，それ以下に存在するコンテンツ数を調査する必要があるが，膨大な処理時間がかかることが予測される．そこで本手法では，既存の WWW 検索システム内に登録されているすべてのコンテンツを WWW 空間全体と置き換えて，WWW 検索システムの URL 検索機能に部分 URL を入力して得られる検索結果数を部分 URL の大域的出現頻度とする．

手順 4：部分 URL の有害度の算出

☆ サーバ内のコンテンツをすべて規制すること

手順1: 部分URLの抽出

http://a2server/home/yoshiyuki/top.html
http://a2server/home/yoshiyuki/adult/index.html
http://a2server/home/issei/main.html

↓

- http://a2server/
- http://a2server/home/
- http://a2server/home/yoshiyuki/
- http://a2server/home/yoshiyuki/top.html
- http://a2server/home/yoshiyuki/adult/
- http://a2server/home/yoshiyuki/adult/index.html
- http://a2server/home/issei/
- http://a2server/home/issei/main.html

手順2: 部分URLの出現頻度の抽出

- http://a2server/ : 3
- http://a2server/home/ : 3
- http://a2server/home/yoshiyuki/ : 2
- http://a2server/home/yoshiyuki/top.html : 1
- http://a2server/home/yoshiyuki/adult/ : 1
- http://a2server/home/yoshiyuki/adult/index.html : 1
- http://a2server/home/issei/ : 1
- http://a2server/home/issei/main.html : 1

手順3: 部分URLの大域的出現頻度の抽出

- http://a2server/ : 400
- http://a2server/home/ : 394
- http://a2server/home/yoshiyuki/ : 8
- http://a2server/home/yoshiyuki/top.html : 1
- http://a2server/home/yoshiyuki/adult/ : 3
- http://a2server/home/yoshiyuki/adult/index.html : 1
- http://a2server/home/issei/ : 2
- http://a2server/home/issei/main.html : 1

手順4: 部分URLの有害度の計算

- http://a2server/ : 0.0075 (3/400)
- http://a2server/home/ : 0.0076 (3/394)
- http://a2server/home/yoshiyuki/ : 0.25 (2/8)
- http://a2server/home/yoshiyuki/top.html : 1 (1/1)
- http://a2server/home/yoshiyuki/adult/ : 0.33 (1/3)
- http://a2server/home/yoshiyuki/adult/index.html : 1 (1/1)
- http://a2server/home/issei/ : 0.5 (1/2)
- http://a2server/home/issei/main.html : 1 (1/1)

図 2 部分 URL ごとの有害度の計算方法
Fig. 2 Calculation method of the hazardous score of partial URL.

手順 2 で求めた部分 URL の出現頻度と手順 3 で求めた大域的出現頻度を用いて，部分 URL の有害度 H_{url} を式 (1) により求める．有害度の大きさは，各部分 URL 以下のコンテンツにどの程度の有害情報があるかを表す割合であるため，0 に近いほど有害性が低く，1 になるほど有害性が高いといえる．部分 URL がデータベース内の末端までの URL と一致する場合，必然的に有害度は 1 となる．

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

3.3 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に対しフィルタリングを行う．Hazardous キーワードによって収集された URL は，すでにデータベース内に登録されているため，確実にフィルタリングすることができる．本手法ではさらに，部分 URL ごとの有害度を判定することで，未登録の URL に対してもフィルタリングを行う．以下にフィルタリング方法の手順を示す．

手順 1：部分 URL 有害度の閾値設定
有害度 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：部分マッチング
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．部分 URL ごとの有害度と閾値 T とを比較し，マッチした部分 URL の有害性の判定を行う．

手順 3：部分 URL の有害性の判定
マッチした部分 URL ごとに 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）と見なし，登録されていなければその URL を無害（Safe）と見なす．

検索結果内の画像にリンクするすべての URL に対して部分マッチングを行い，Hazardous と見なされた URL にリンクを張っている画像を制限することで，検索結果内の有害画像をフィルタリングすることができる．部分マッチングによる本フィルタリング手法の例を**図 3** に示し，以下で各閾値の場合ついて説明する．

- 閾値 $T = 0.3$ の場合
  Pattern 1，2，3 の有害度はすべて閾値未満になるため，その URL は Safe と見なされる．また，Pattern 3 以降のパスには，マッチする URL が存在しないため，入力 URL は Safe URL として判定される．
- 閾値 $T = 0.22$ の場合
  Pattern 1，2 は Safe と見なされるが，Pattern 3 が閾値以上となるため，入力 URL は Hazardous URL として判定される．
- 閾値 $T = 0.00755$ の場合
  Pattern 1 は Safe と見なされるが，Pattern 2 が

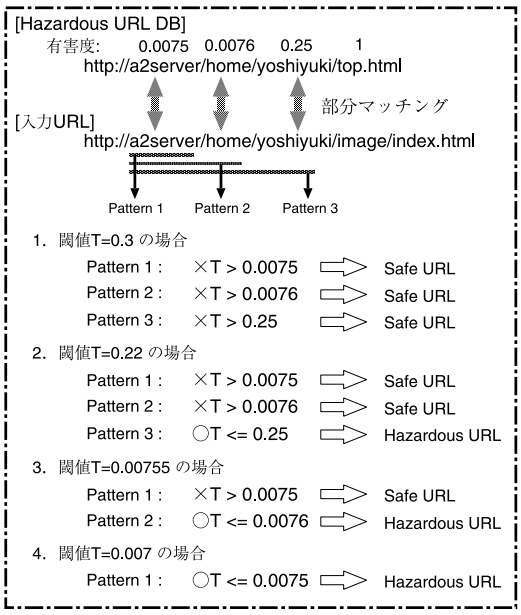

図 3 部分マッチングによる有害な URL の判定
Fig. 3 Estimation hazardous score using partial URL-based weighing scheme.

閾値以上となるため，入力 URL は Hazardous URL として判定される．

- 閾値 $T = 0.007$ の場合
  Pattern 1 で閾値以上となるため，入力 URL は Hazardous URL として判定される．

4. Hazardous キーワードの選定

4.1 Hazardous URL DB 内に混入する Safe URL による影響

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 に登録する．しかし，この URL の中には有害な画像を含んでいない Safe な URL 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．実際に，本手法により自動構築した Hazardous URL DB から人手で Safe な URL を除去して比較したところ，Hazardous 画像の正解率に平均 10%の向上が見られた．

また，キーワードごとに Safe な URL が混入する割合に大きな差があることが分かった．これは，ユーザが設定した Hazardous キーワードに有害性が低い単語や意味的多義性を有する単語が含まれていたことが原因であった．このような意味的多義性を含む単語をユーザが意識せずに Hazardous キーワードとして使用すると，フィルタリング精度を低下させてしまう．たとえば，「処女」という単語は「処女作」や「処女航海」といった使い方もあり，Hazardous キーワードとしては不適切である．

このことから，データベース構築の際に有害性の高い Hazardous キーワードのみを使用すれば，さらに高精度なフィルタリングができると考えられる．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．

4.2 有効な Hazardous キーワードの選定手法

本選定手法は，Hazardous キーワードごとに検索結果内の画像を含むページまたはリンクするページ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．具体的にはユーザが設定した Hazardous キーワードのほとんどが適切だと仮定すると，Hazardous キーワードで検索された各コンテンツ内で，検索で用いたキーワードとは別の Hazardous キーワードの種別数（Hazardous キーワードの異なり数）を算出することにより，その値が低いキーワードを有害性が低いと判断する．たとえば，「処女」といった単語が Hazardous キーワードに含まれていた場合，「処女作」や「処女航海」といった単語が存在するページには，他の Hazardous キーワードはほとんど出現しないため，異なり数が低くなり除外の対象となる．この共起性に従った選定手法の手順を以下に示す．

手順 1：画像にリンクするページを取得
Hazardous キーワード $key_i (1 \leq i \leq n)$ ごとに検索結果内の画像にリンクするページ $HTML_{ij} (1 \leq j \leq m)$ を取得する．

手順 2：異なり数の算出
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 $H(key_i)$ を求める．

手順 3：Hazardous キーワードの選定
求めた平均値を Hazardous キーワードの有害度とし，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．

本選定手法の流れを**図 4** に示す．この手法を用いることにより，Safe が多く混入している可能性が高い Hazardous キーワードを自動で除外することができる．

5. 評価実験

本手法の有効性を確認するため，URL の部分マッチングによるフィルタリング手法および Hazardous キーワード選定手法を用いて，既存の WWW 画像検

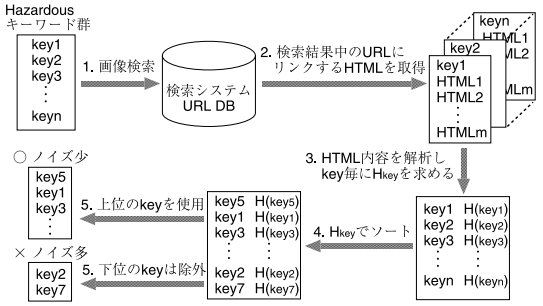

図 4 Hazardous キーワード選定手法
Fig. 4 Selection method of hazardous keywords.

索システムの検索結果に対するフィルタリング実験を行った．以下に実験条件，評価基準，各手法に対する実験結果，考察を述べる．

5.1 実験条件

既存の WWW 画像検索システムに Google Image Search を用いて，Hazardous URL DB と評価用データを作成した．まず，k 個の Hazardous キーワード☆で検索し，検索結果上位 100 件の画像ページの URL をデータベースに登録した．k は各評価実験により異なるため，評価ごとの実験方法で説明する．次に，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「女子高生」や「制服」といった 27 個の評価用キーワード☆☆で検索を行い，検索結果上位 100 件の画像ページの URL 計 2,639 件を評価用データとした．さらに，評価用データ中の URL 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 と 2 種類に分類し，456 件の Hazardous URL と 2,183 件の Safe URL を得た．

5.2 評価基準

フィルタリング精度の評価尺度には，再現率・適合率および F 尺度を用いた．評価用データに対して Hazardous URL DB を用いてフィルタリングを行い，式 (2)，(3) に示す 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．

$$R_{haz} = \frac{\text{正しく } Hazardous \text{ と分類された画像数}}{\text{全 } Hazardous \text{ 画像数}} \quad (2)$$

$$P_{haz} = \frac{\text{正しく } Hazardous \text{ と分類された画像数}}{Hazardous \text{ と分類された画像数}} \quad (3)$$

また，Hazardous 画像の再現率・適合率を求めると同時に，式 (4)，(5) に示す Safe 画像の再現率（R_{saf}），適合率（P_{saf}）もあわせて求めた．R_{saf} は評価用データ中の全 Safe 画像に対してアクセスを許す割合を表し，P_{saf} はアクセスを許した画像の中で本当に Safe 画像であった割合を表す．

$$R_{saf} = \frac{\text{正しく } Safe \text{ と分類された画像数}}{\text{全 } Safe \text{ 画像数}} \quad (4)$$

$$P_{saf} = \frac{\text{正しく } Safe \text{ と分類された画像数}}{Safe \text{ と分類された画像数}} \quad (5)$$

本手法では部分 URL の有害度に閾値を設定してフィルタリングを行うため，閾値ごとに再現率・適合率は変化する．そこで，閾値 T を 0.0〜1.0 まで 0.0001 ごとに変化させ，それにより得られた再現率・適合率をプロットし，再現率・適合率曲線[14] を求めた．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．F 尺度は以下の式 (6) で求めることができる[14]．

$$F = \frac{2RP}{R+P} \quad (6)$$

再現率・適合率曲線ではグラフ中に多数の再現率と適合率のペアが存在するため，各曲線において，再現率を 0.0〜1.0 まで 0.05 ごとに区切った計 101 点の F 尺度を計算し，その平均値を求めた．

5.3 URL の部分マッチングによる評価

5.3.1 実験方法

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 Hazardous URL DB に登録した．次に以下の 4 つ手法を用い，評価用データに対してフィルタリングを行い，再現率・適合率および F 尺度を求めた．

- All
  本手法による，部分 URL ごとの有害度を用いた重みづけ．
- Normal
  部分 URL の大域的出現頻度を用いた正規化を施さず，Hazardous URL DB 中の部分 URL の出現頻度のみを用いた重みづけ．Hazardous キーワードで収集された URL を多く持つサーバであるほど有害情報を多く含んでいる．そのため従来のブ

☆ 検索キーに対し制限を行う機能を提供している goo を用いて制限されるキーワードを調査し，そのキーワードを用いて Google Image Search で検索を行い上位 100 件の検索結果内に有害画像が 7 割以上を占めた検索キーワードを選択した（2005 年 3 月現在）．

☆☆ goo で制限されないキーワードを用い Google Image Search で検索を行い，有害画像が検索されるキーワードを選択した．

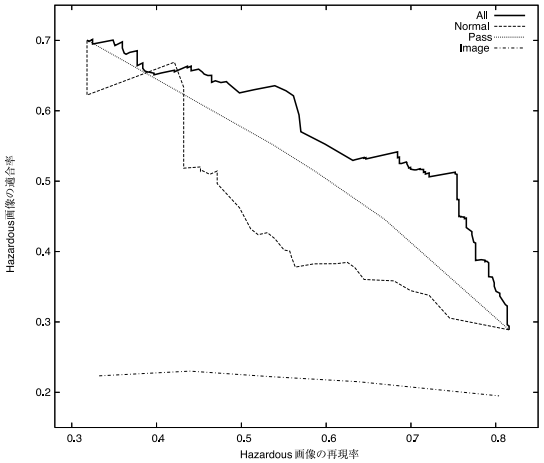

図 5 Hazardous 画像の再現率・適合率
Fig. 5 Recall-precision curve of hazardous image.

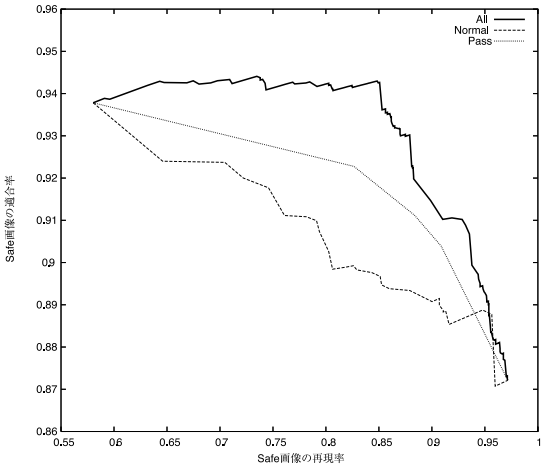

図 6 Safe 画像の再現率・適合率
Fig. 6 Recall-precision curve of safe image.

ラックリスト方式で問題となっているサーバ全体の包括規制に該当する．

- Pass
  部分 URL の出現頻度を考慮せずパスの深さのみを用いた重みづけ．URL のパス数を d とすると，URL のサーバ部を深さ 1，サーバ以下の第 1 パス部を深さ 2，第 2 パス部を深さ 3，...，ファイル部を深さ d として重みづけを行う．
- Image
  既存の WWW 画像検索システムを用いて，人手で Hazardous 画像を 500 件抽出し，色情報，形状情報を用いて画像解析により Hazardous 画像を自動で判定．距離計算方法には画像検索の分野で高精度な検索を実現する Earth Mover's Distance [15] を用いた．この手法は，コンテンツチェック方式に該当する．

5.3.2 実験結果

Hazardous 画像の再現率・適合率曲線を**図 5** に，Safe 画像の再現率・適合率曲線を**図 6** に示す．また各手法の F 尺度の平均値を**表 1** に示す．

Normal に比べ 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．これは，従来のブラックリスト方式で問題であった包括規制を緩和できているといえる．また，Pass に比べ All が高い値を示していることから，Pass ではサーバやディレクトリで過剰な規制が行われてしまうのに対し，All では過剰な規制を防ぎつつ URL の有害性を部分的に識別できていると考えられる．Image では再現率の値にかかわらず適合率が低い値となった．これは肌色の画像をすべて Hazardous 画像と見なすため，適合率が悪くなったと考えられる．

表 1 各手法における F 尺度の平均値
Table 1 Average F measure for each method.

	All	Normal	Pass	Image
haz	0.6955	0.6522	0.6699	0.2857
saf	0.8469	0.8243	0.8648	0.3922

5.4 Hazardous キーワードの選定による評価

5.4.1 実験方法

54 個の Hazardous キーワードの中から，有害度の高い上位 10 件，20 件，30 件，40 件のキーワード選定を行った．この 4 通りに選定したキーワードセットおよび 54 個のキーワードセットを用いて，Google で検索し，検索結果上位 100 件の URL をデータベースに登録し，本提案手法により重みづけを行った．各キーワードセットで構築した URL 数を以下に示す．

- 上位 10 件による選定（key10）：752 件
- 上位 20 件による選定（key20）：1,544 件
- 上位 30 件による選定（key30）：2,310 件
- 上位 40 件による選定（key40）：3,061 件
- 選定なし（key54）：4,189 件

5 つのデータベースを用い，評価用データに対してフィルタリングを行い，再現率・適合率を求める．

5.4.2 実験結果

Hazardous 画像の再現率・適合率曲線を**図 7** に，Safe 画像の再現率・適合率曲線を**図 8** に示す．

まず選定処理に対する考察を行う．実験結果より，上位 40 件に選定したキーワードセット（key40）は選定なしのキーワードセット（key54）に比べ Hazardous 画像の再現率を保ちつつ適合率が向上しており，選定処理の有効性が確認できる．これは Hazardous キー

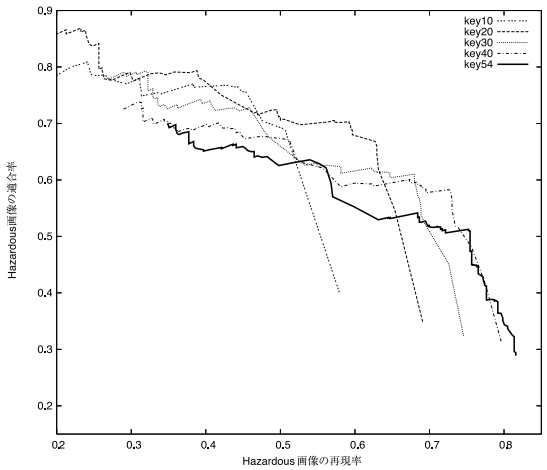

図 7 選定された Hazardous キーワード数別の Hazardous 画像の再現率・適合率

Fig. 7 Recall-precision curve of hazardous image for each number of selected hazardous keywords.

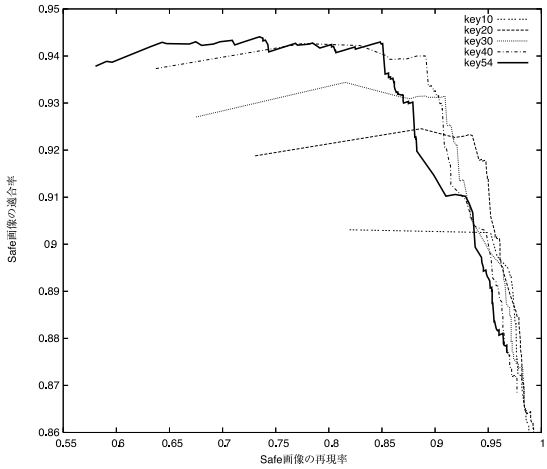

図 8 選定された Hazardous キーワード数別の Safe 画像の再現率・適合率

Fig. 8 Recall-precision curve of safe image for each number of selected hazardous keywords.

ワードの選定を行うことにより，Hazardous URL DB に混入する Safe URL の割合が少なくなっているためである．この結果より，Hazardous キーワードの選択基準を現状の 7 割から緩めると，Safe URL の割合が多くなるため，Hazardous 画像の適合率が低下することが予測できる．逆に選択基準を厳しくすると，適合率は向上するがデータベース構築のために採用できる Hazardsou キーワード数が減少してしまうことが予測できる．

次に選定後の結果に着目し，Hazardous キーワード数の変化に対する考察を行う．実験結果より，キーワード数が少なくなればなるほど Safe URL の混入する割合が少なくなるので，Hazardous 画像の適合率は向上する．しかし，キーワード数を少なくしすぎると Hazardous 画像の再現率が顕著に低下するため，キーワード数を絞り込みすぎることはフィルタリング処理において大きな問題を生じることになる．

以上のことより，有害画像フィルタリングシステムの本来の目的である，Hazardous 画像の適合率を保ちつつ再現率を 100%に近づけるためには，キーワードの選定処理を施した Hazardous キーワードをできるだけ多く準備して，Hazardous URL DB を構築する必要がある．そのためには，数十件程度の有効な Hazardous キーワードを人手で設定し，関連キーワードの自動収集手法[16] などを用いて Hazardous キーワードの拡張を行いつつキーワードの選定をすれば，高精度なフィルタリングができると考えられる．

6. ま と め

本論文では数十件の Hazardous キーワードを準備するだけで，既存の WWW 画像検索システムの検索結果から有害画像をフィルタリングするのに有効な URL データベースの構築手法，および URL をパスごとに重みづけし，有害性の高い URL を部分的に識別することでフィルタリングする手法を提案した．また，URL データベースの自動構築時に問題となるノイズ混入を防ぐため，Hazardous キーワードの意味的多義性に着目したキーワードの選定手法を提案した．評価実験では従来の方式に比べフィルタリング精度を向上させることができた．また，URL データベースに混入するノイズを除去することに成功した．

今後は，関連キーワードの自動収集手法[16] を用いて Hazardous キーワードの拡張を行い，拡張したキーワード群からさらにキーワードの選定を行えば，より有効な URL データベースが構築でき，高精度なフィルタリングができると考えられる．

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) 国分明男，清水 昇：インターネットにおけるコンテンツ・レイティングとフィルタリング，情報処理学会論文誌，Vol.40, No.1, pp.57–63 (1999).
2) 井ノ上直己，帆足啓一郎，橋本和夫：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，電子情報通信学会論文誌，Vol.J83-DII, No.6, pp.1158–1166 (2001).
3) Oder, N. and Kenney, B.: CIPA Fallout: ALA

Cancels Meeting with Filter Makers, *Library Journal*, Vol.128, No.15, pp.14–15 (2003).
4) Google Homepage. http://www.google.co.jp/
5) goo Homepage. http://www.goo.ne.jp/
6) AltaVista Homepage. http://www.altavista.com/
7) Yahoo Homepage. http://www.yahoo.co.jp/
8) Calishain, T., Dornfest, R.（著），山名早人，田中裕子（訳）：*GOOGLE HACKS*, オライリー・ジャパン (2003).
9) 文部省・郵政省：子どもたちがもっと自由にインターネットを活用できる環境づくりを目指して，教育分野におけるインターネットの活用促進に関する懇談会 (1998).
10) 榎本　聡，室田真男，清水康敬：漢字かな自動変換機能等を備えたインターネット学習システムの開発，電子情報通信学会論文誌，No.3, pp.384–394 (2000).
11) 山名早人：情報検索の新潮流，*Computer Today*, No.87 (1998).
12) W3C Homepage. http://www.w3.org/PICS/
13) 財団法人ニューメディア開発協会 Homepage. http://www.iajapan.org/rating/
14) 北　研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版 (2002).
15) Rubner, Y., Guibas, L.J. and Tomasi, C.: The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval, *Proc. ARPA Image Understanding Workshop*, pp.661–668 (1997).
16) 山本一晴，獅々堀正幹，柘植　覚，北　研二：出現 URL の類似性に着目した WWW 空間からの関連語自動収集手法，情報処理学会自然言語処理研究会資料，NL-170, pp.45–52 (2005).

(平成 17 年 12 月 20 日受付)
(平成 18 年 4 月 10 日採録)

(担当編集委員　飯沢 篤志)

小泉　大地（学生会員）

平成 14 年徳島大学工学部知能情報工学科卒業．平成 16 年同大学大学院工学研究科博士前期課程知能情報工学専攻修了．現在同大学院工学研究科博士後期課程情報システム工学専攻 2 年．マルチメディア情報検索の研究に従事．

獅々堀正幹（正会員）

平成 3 年徳島大学工学部情報工学科卒業．平成 5 年同大学大学院博士前期課程修了．平成 7 年同大学院博士後期課程退学．同年徳島大学工学部知能情報工学科助手．平成 9 年同講師．平成 13 年同助教授．博士（工学）．マルチメディア情報検索の研究に従事．著書『情報検索アルゴリズム』（共立出版）．情報処理学会第 45 回全国大会奨励賞受賞．

中川　嘉之

平成 15 年徳島大学工学部知能情報工学科卒業．平成 17 年同大学大学院工学研究科博士前期課程知能情報工学専攻修了．同年インフォコム株式会社入社．

柘植　　覚（正会員）

平成 8 年徳島大学工学部知能情報工学科卒業．平成 10 年同大学大学院工学研究科博士前期課程知能情報工学専攻修了．平成 13 年同大学大学院工学研究科博士後期課程システム工学専攻修了．平成 12 年徳島大学工学部助手，博士（工学）．音声認識，情報検索等の研究に従事．日本音響学会会員．

北　　研二（正会員）

昭和 56 年早稲田大学理工学部数学科卒業．昭和 58 年沖電気工業（株）入社．昭和 62 年 ATR 自動翻訳電話研究所出向．平成 4 年徳島大学工学部講師．平成 5 年同助教授．平成 12 年同教授．平成 14 年同大学高度情報化基盤センター教授．博士（工学）．自然言語処理，情報検索等の研究に従事．平成 6 年日本音響学会技術開発賞受賞．著書：『確率的言語モデル』（東京大学出版会），『情報検索アルゴリズム』（共立出版）等．電子情報通信学会会員，言語処理学会会員．