# WWW有害情報のフィルタリングのための画像判別手法
# Image discriminant method for filtering web pages with potentially harmful information

I-82

武者 義則† 広池 敦† 森本 康嗣† 松田純一†
Yoshinori Musha Atsushi Hiroike Yasutsugu Morimoto Jyunichi Matsuda

## 1. まえがき

近年，インターネットの普及に伴い，年々急速にホームページ数が増大している．また，ホームページの作成・公開が匿名で可能なため，悪意をもったユーザが公開した不適正な内容でさえ，子供でも容易にアクセスできてしまう．一方，フィルタリングシステムが言論の自由を不当に制限してはならないという問題がある．そこで，W3C は情報発信を制限せず，第三者のレイティング情報に基づいて受信者が受信内容をコントロールする枠組みとして PICS を提案した[1]．しかし，レイティングは人的負荷が高く，web ページ数が".jp"ドメインだけで年々1500～2000 万のペースで増大する状況の下では，その人的負荷の軽減や web ページの自動分類の技術が望まれている．今日，キーワードやテキスト情報を用いた自動フィルタリングシステムが実際に運用されている[2]．

我々は，テキスト情報と画像情報を用いて web ページを分類するシステムを開発した[4]．本システムはレイティング情報のない有害ページを自動的にフィルタリングし，またオペレータによる web ページの格付け作業を効果的に支援することができる．本稿では，本トータルシステムの構成と，特に web ページのための画像分類手法について報告する．また，フィルタリングと格付け支援の両ケースについて，画像判別に基づいた実験結果を示す．

## 2. フィルタリングのためのトータルシステム

本トータルシステムは，URL ベースのフィルタリングシステム，内容ベースのフィルタリングシステム，レイティング支援システムの３つで構成されている．図１にデータフローを示した．小さい子供の親を含むユーザは，あらかじめカテゴリ毎に望むレベルを設定しておく．RSACi レイティング基準を元に日本で拡張された SafetyOnline レイティング基準に基づき，５つのカテゴリ，(n)，(s)，(v)，(l)，(e)，および各カテゴリに対して５段階（0～4）のレベル基準が用意されている[3]．その基準において，数の小さいレベルは，厳しいフィルタリングで子供を含むすべてのユーザに受け入れられるコンテンツを意味し，数の大きいレベルは緩いフィルタリングを意味する．ユーザが web ページにアクセスした時，まず URL ベースのフィルタリングシステムがその URL をチェックし，その格付け情報がユーザの登録したレベルより低ければユーザはアクセスできる．それ以外であればアクセスは拒否される．ただし，格付け情報がなければ，内容ベースのフィルタリングシステムがテキストと画像をチェックし，ユーザが登録したレベルにおける適不適の判定を行う．その際の web ページは URL 格付けデータベースへ保存され，格付け推定システムで推定された結果が格付け作業者へ示される．格付け作業者は，格付け属性を web ページに付与してデータベースへ格納する．その属性情報は URL ベースのフィルタリングに使われると同時に，内容ベースフィルタリングや格付け推定の精度向上に使われる．

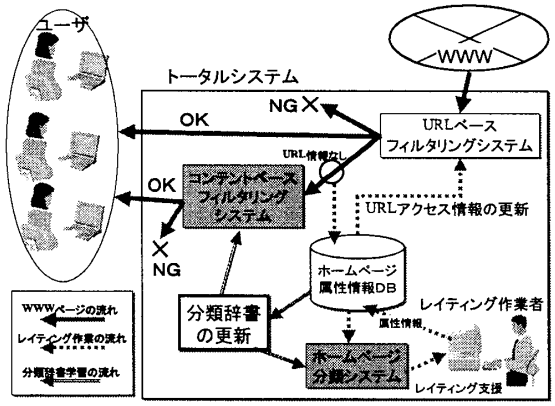


図1 トータルシステム

## 3. 画像判別によるフィルタリングおよびレイティング支援

画像判別による内容ベースのフィルタリングシステムおよび格付け推定システムの構成を図 2，図 3 に示した．両者とも同じ４つの(n)カテゴリのレベル境界判別システムと４つの(s)カテゴリのレベル境界判別システムを備えている．それぞれの判別システムは 200 次元の画像特徴量を用い，入力画像の特徴量と辞書中のラベル付けされた特徴量を比較することによって判別を行う．図 2 のフィルタリングシステムにおいて，(1)HTML ファイルに貼付された画像のうち 64x64 以上のサイズのみ選別される．(2)次に，ユーザが指定したレベル境界判別システムで画像が判別され，(3)その全ての画像の判別結果が適正の場合のみ HTML 単位の判定が適正となるよう統合される．これにより，HTML 単位の適正フィルタリングの精度が画像単位の精度よりも高くなり，不適正の再現率も高くなることが期待される．一方，図 3 の格付け推定システムの場合，前述(1)のサイズで選別された全画像が，全てのレベル境界判別システムによって判別が行われ，ベイズ推定によっ

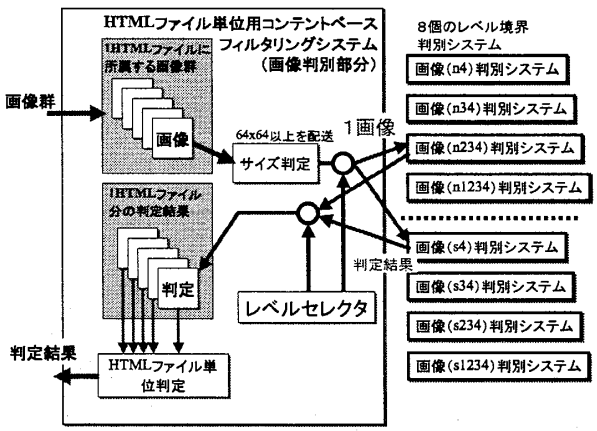


図2 内容ベースのフィルタリングシステム

† （株）日立製作所中央研究所
マルチメディアシステム研究部

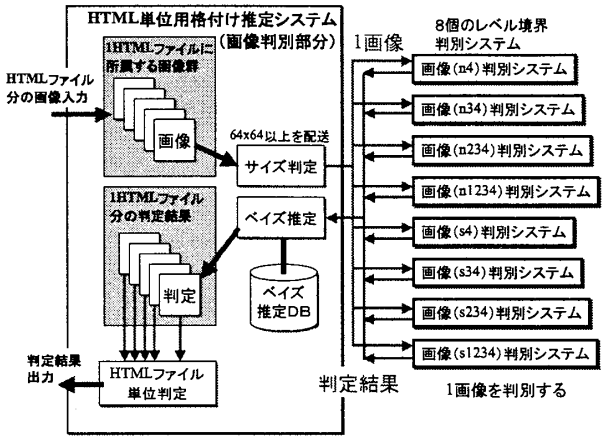


図3　レイティング支援システム

て各レベルの出現確率が計算される．我々はその確率をそのレベルの信頼度として扱う．そして，その全ての画像の最も高いレベルを HTML ファイルのレベルとするように，HTML ファイル単位で結果が統合される．

## 4．フィルタリングとレイティング推定の評価

我々は，画像判別に基づいた(A)フィルタリングと(B)格付け推定の実験を行った．

(A) HTML ファイルに貼付された 6,383 像を用意し，(n)カテゴリについて５つのレベル[3]を付与した．更に，画像を HTML 単位で２セットに分け，一方を学習セット，他方を評価セットとして用いた．図４は画像単位と HTML 単位の精度を，図５は再現率を示す．例えば，(n234)判別システムは，n0, n1 レベルを適正とみなし，n2, n3, n4 レベルを不適正とみなす．全ての判別において，適正検出の精度と不適正検出の再現率で，HTML 単位の方が高くなった．

(B) HTML ファイルに貼付された 6,531 画像を用意し，(n)カテゴリについて５つのレベルを付与した．更に，画像を HTML 単位

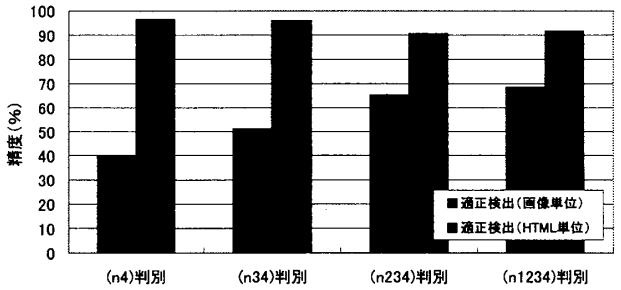


図4　画像判別における適正検出（精度）

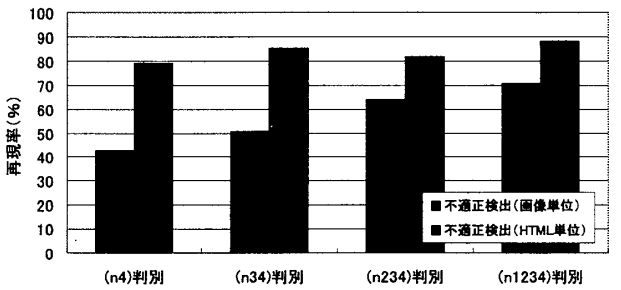


図5　画像判別における不適正検出（再現率）

表2　格付け推定における精度

| | 画像単位 | | HTML単位 | | レベル候補の閾値 |
|---|---|---|---|---|---|
| | 精度(%) | アクセプト(%) | 精度(%) | アクセプト(%) | |
| n0 | 83.49 | 12.8 | 92.33 | 12.73 | 0.811 |
| n1 | 11.24 | 13.15 | 10.15 | 12.11 | 0.095 |
| n2 | 21.98 | 15.19 | 19.01 | 10.27 | 0.22 |
| n3 | 20.42 | 13.65 | 10.26 | 11.28 | 0.19 |
| n4 | 32.28 | 14.54 | 35.67 | 25.74 | 0.093 |
| all | 33.15 | Reject=30.68 | 35.04 | Reject=0.00 | --- |

で学習セット，ベイズ学習セット，評価セットの３つに分け，これらの組み合わせで６回実験を行った．また，格付け推定の際に，精度向上を期待してリジェクト機能を使用した．各レベルにおいて相対的に低い信頼度を持つレベル候補をリジェクトする．また，各レベルにおいてアクセプト率が 16%以下となるようにリジェクトの閾値を設定した．実際には画像全体のリジェクト率は 30.68%となった．全レベルの中では n0 が最も高い精度となった．また，HTML 単位の精度の方が画像単位よりも高いことを期待していたが，レベル n0, n4 のみでその傾向が見られた．ほかの n1, n2, n3 において，HTML 単位の精度が低くなった理由として，アクセプト率の低下が考えられる．

## 5．おわりに

我々は，web ページをフィルタリングするトータルプロトタイプシステムを開発した．そのシステムは，web ページをフィルタリングするとともに，格付け推定により，格付け作業者を支援することができる．また，フィルタリングシステムと格付け推定システムの両方に，同じレベル境界判別システムを用いているため，両者の精度を同時に向上させることができる．特に，本システムは，ユーザが一度アクセスした web ページに類似したページに関し精度が向上するシステム構成となっている．更に，我々は画像判別に基づいたフィルタリングと格付け推定の実験を行った．本システムでは，HTML に貼付された画像群に関して，HTML 単位で評価することにより適正検出の精度と不適正検出の再現率を向上させることができた．フィルタリングにおいて適正検出がよい傾向にあるのは子供などを有害情報に触れさせたくないという SafetyOnline レイティング基準の主旨と合致している．とりわけ，WWW 上に大量に氾濫し子供へ直接影響を与えてしまう不適正な画像情報を，画像自体によってフィルタリングすることは非常に重要である．しかし，格付け支援を効果的に行うためには，更に精度向上を行う必要がある．なお，本研究は総務省の認可法人である通信・放送機構（TAO）の委託研究[5]として行われた．

## 参考文献

1) Internet Content Rating Association: http://www.icra.org/
2) 井ノ上, 帆足, 橋本: 文章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.
3) 財団法人ニューメディア開発協会: http://www.nmda.or.jp/enc/rating/eindex.html
4) 武者, 広池: WWW 有害情報のレイティング支援のための画像判別手法, 信学会総合大会, D-12-33, p.209, 3 月, 2002.
5) 松田, 本城: 情報通信不適正利用対策技術の研究開発, 通信放送機構（TAO） 平成 14 年度研究発表会予稿集 <席上発表編>, pp. 45—49, 5 月, 2002.