RD-002

HTML 要素に着目した違法・有害サイト検出手法の提案と評価
Detection of Illegal and Hazardous Information Based on HTML Elements

池田 和史　柳原 正　松本 一則　滝嶋 康弘†
Kazushi Ikeda　Tadashi Yanagihara　Kazunori Matsumoto　Yasuhiro Takishima

1. まえがき

インターネットの普及により、一般ユーザ向けの Web サイトや掲示板が増加している。出会い系サイトや犯罪予告サイト、誹謗・中傷などの書き込みを含む学校裏サイトなど、違法・有害な情報を含むサイトも増加傾向にあり、目視によるサイトの監視に要するコストは大きなものとなっている。近年、違法・有害な Web サイトを自動的に検出するためのフィルタリングシステムの開発が進んでおり、ウェブブラウザに組み込まれてリアルタイムに違法・有害サイトを検出したり、Web サイトの監視事業者が膨大な Web サイトの中から違法・有害性の高いサイトを優先的に目視により監視するなどの利用シーンが想定されるため、高精度かつ高速な判定が可能な違法・有害サイト検出手法が求められる。

既存の主流な違法・有害サイト検出手法として Web サイトの URL を利用する Black/White リスト方式があるが、データベースを管理する人的コストが大きい点や、ブログなどでは同一ドメイン下に違法・有害サイトと無害サイトの両方が存在するために判定精度が低下する点、新規のサイトに対して判定が行えない点などが課題として挙げられる。これに対し、Web サイトに記載の文書や掲載された画像を解析し、文書に特定のキーワードが含まれていることや画像の特徴を利用することで、違法・有害サイトを検出するコンテンツベースの手法も提案されているが、単純な方式では高精度に違法・有害サイトを検出することは難しく、一方で高度な言語処理や画像処理を行う手法では処理時間が大きくなることが課題である。

総務省が 2008 年に実施した調査[1]によると、インターネット上で公開されている国内のブログの総数は 1690 万ブログ(記事総数は 13 億 5000 万記事)存在し、毎月 4000 万記事が新規に投稿されると言われる。違法・有害な記事の割合はブログの運営事業者によっても異なるが、例えば全体の 10%が違法・有害な記事であると仮定し、監視事業者が違法・有害な 400 万記事のうち 280 万記事を発見、削除するというタスクを考える(再現率は 70%となる)。フィルタリングシステムの適合率は一般に 100%に満たないため、監視事業者は無害な記事を誤って違法・有害と判定して削除しないように、最終的には人手で目視を行った後に記事を削除するが、このときフィルタリングシステムによって違法・有害性が高いと判定された記事から優先的に目視を行うことで、作業を効率化するものと想定される。ここで、フィルタリングシステムの適合率が 60%の場合、違法・有害な 280 万記事を発見するには 280 万 / 60% = 467 万 記事を目視により確認する必要がある(すなわち無害な記事を 187 万記事確認することになる)が、適合率が 70%の場合、280 万 / 70% = 400 万 記事の確認により目標を達成できる(無害な記事は 120 万記事しか確認せずに済む)。目視可能な記事数を 1 万記事/人日とすると、削減可能な人的コストは大きい。また、フィルタリングシステムにおける処理時間の短縮も運営設備の削減などコストの削減につながる。

本稿では高速かつ高精度に違法・有害サイトを検出するため、Web サイトの HTML を対象とした違法・有害サイト検出手法を提案する。提案手法では違法・有害サイトの HTML に偏って出現するような文字列を自動的に抽出し、SVM(Support Vector Machine)を用いてこれらの特徴を組み合わせて違法・有害サイトの検出を行う。提案手法は Web サイトの本文の情報を利用しないため、既存のキーワードベース方式によって検出が困難なサイトも検出が可能である点が特徴である。このため、既存のキーワードベース方式と組み合わせて利用することも有効である。

性能評価実験においては、人手によって違法・有害または無害のラベルが付与された学習用 Web サイトと判定対象 Web サイト各 2 万サイトを利用した大規模な実験を実施した。提案手法を単独で利用した場合で再現率 50.0%、適合率 90.3%など極めて高い適合率が実現できることを確認した。加えて、既存のキーワードベース方式と提案手法を組み合わせて判定を行う複合手法では再現率 70.0%、適合率 78.1%となった。これはキーワードベース方式を単独で利用した場合の同程度の再現率と比べて適合率が 9.3%向上しており、極めて高性能なフィルタリングシステムを実現したと言える。

2. 関連研究

Web サイトに記載の文書情報を利用して違法・有害サイトを自動的に検出するいくつかの手法が提案されている[2],[3]。文献[2]の手法では、学習用文書において違法・有害な文書に偏って出現する単語を違法・有害キーワードとして情報量基準に基づき統計的に抽出し、キーワードが判定対象文書に含まれていれば違法・有害として検出する。形態素解析を用いることなく判定が可能であり、判定ロジックも単純なキーワードマッチングであるため処理は高速であるが、精度に課題がある。複数のキーワードを組み合わせた判定や係り受け解析などを用いた深い言語解析を行うことで高精度化が可能となるが、高度な言語処理は処理時間が大きくなる。文献[3]の手法では、学習用文書と判定対象文書の特徴ベクトルをそれぞれ求め、判定対象文書の特徴ベクトルが学習用の違法・有害文書の特徴ベクトルとどの程度類似しているかによって、判定対象文書の違法・有害度合いを算出する。この手法では、判定対象文書に対して形態素解析を行う必要があるため、処理時間が大きくなることが課題である。

Web サイトの画像数やリンク数といった HTML に関連する特徴を用いて Web サイトの分類を行う手法も提案されている[4],[5]。文献[4]では、人手により Web サイトを観測することで、違法・有害サイトの判定に役立つと思われる特徴を発見し、判定に利用する手法が提案されている。文献[5]も同様に、違法・有害サイトの検出に役立つ特徴として画像数やリンク数などを挙げ、リンク数が 10 以上の

† （株） KDDI 研究所, KDDI R&D Laboratories Inc.

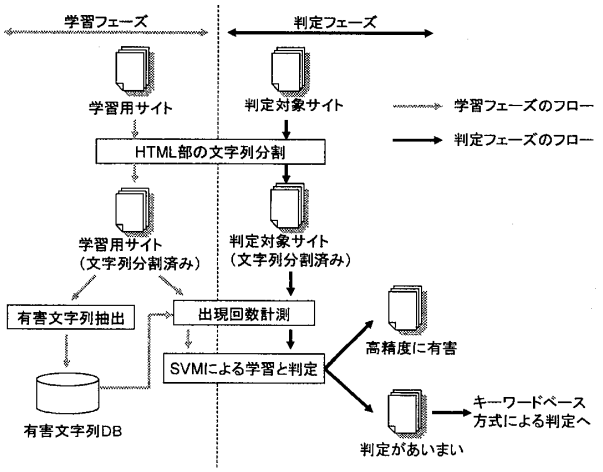

図 1　提案手法における処理フロー

サイトは無害サイトに比べ違法・有害サイトの方が多い、といった傾向を発見し、それらの特徴を組み合わせてベイジアンネットワークを用いて判定に利用している。しかし、これらの手法で抽出可能な特徴は観測者の主観や閲覧した Web サイトに依存するため、十分な性能を得ることが難しい。例えば、著者らの予備実験において、違法・有害サイトおよび無害サイト各 1 万サイトに対し、リンク数が 10 以上のサイトを全て違法・有害と判定したとすると、違法・有害サイト全体の 75.7%を検出することができた(再現率=75.7%を意味する)が、違法・有害と判定したサイトのうち、実際に違法・有害であったサイトは 56.8%であり(適当率=56.8%を意味する)、特徴量としての有効性は低いと考えられる。

Web サイトのハイパーリンクやソーシャルネットワークサービスの知り合い関係などを用いて Web サイトの分類を行う研究も報告されている[6],[7]。文献[6]では、ハイパーリンクの共起性とベクトル空間モデルを用いたクラスタを重ね合わせることにより、類似したクラスタを検出し、分類を実現している。文献[7]では、社会ネットワーク分析で用いられる指標を利用し、リンクに基づいてノードを高精度に分類する手法が提案されている。本稿では Web サイト単体で判定を行う手法を提案しているが、これらの文献の知見を応用することで、さらなる高精度化が可能であると考えられる。

3. 提案手法

3.1. 提案手法の概要

提案手法における違法・有害サイト検出処理の概要を図 1 に示す。提案手法では、違法・有害または無害のラベルが人手により付与された学習用サイトを利用した学習フェーズと判定対象となるサイト集合から違法・有害なサイトを検出する判定フェーズがある。学習フェーズでは始めに、違法・有害サイトの HTML に偏って出現するような文字列を統計的な基準を用いて自動的に抽出する。次に、抽出した違法・有害性の高い各文字列の学習用サイトにおける出現回数を特徴量として、SVM の学習を行う。判定フェーズでは学習フェーズと同様に、違法・有害性の高い各文字列の判定対象サイトにおける出現回数を特徴量として SVM を用いて判定を行う。

表 1　HTML 要素の抽出と文字列分割の例

HTMLソース

```
<td><img src="img/gaiyo.gif" width="560" height="25" alt="開催概要" /></td> </tr> <tr> <td height="80" valign="top" class="font_glay_11">電子情報通信学会情報・システムソサイエティ(ISS)及びヒューマンコミュニケーショングループ(HCG)と情報処理学会(IPSJ)の合同で開催致します本フォーラムは，IPSJ全国大会とISSソサイエティ大会との流れを汲むものですが，従来の大会の形式にとらわれずに，新しい発表形式を導入し，タイムリーな情報発信，活気ある議論・討論，多彩な企画，他分野研究者との交流などを実現してゆきたいと考えております．<br /> ※<a href="http://www.ipsj.or.jp/10jigyo/fit/fit_found.html" target="_blank">FIT創設の経緯とIPSJ-ISS覚書</a></td>
```

本文テキスト

```
電子情報通信学会情報・システムソサイエティ(ISS)及びヒューマンコミュニケーショングループ(HCG)と情報処理学会(IPSJ)の合同で開催致します本フォーラムは，IPSJ全国大会とISSソサイエティ大会との流れを汲むものですが，従来の大会の形式にとらわれずに，新しい発表形式を導入し，タイムリーな情報発信，活気ある議論・討論，多彩な企画，他分野研究者との交流などを実現してゆきたいと考えております．※FIT創設の経緯とIPSJ-ISS覚書
```

本文テキストを除いたHTML要素

```
<td><img src="img/gaiyo.gif" width="560" height="25" alt="開催概要" /></td> </tr> <tr> <td height="80" valign="top" class="font_glay_11"><br /> ※<a href="http://www.ipsj.or.jp/10jigyo/fit/fit_found.html" target="_blank"></a></td>
```

HTML要素を分割した文字列(括弧内は複数出現回数)

```
a(2), alt, blank, br, class, fit(2), font, found, gaiyo, gif, glay, height(2), href, html, http, img(2), ipsj, jigyo, jp, or, src, target, td(4), top, tr(2), valign, width, www
```

提案手法は Web サイトの HTML 部分のみを用いて判定を行うため、本文を対象として判定を行う既存のキーワードベース方式と組み合わせて利用することでさらに高精度な判定を行うことが可能と考えられる。3.5 節では、提案手法と既存のキーワードベース方式[2]によって検出可能な違法・有害サイトの相関関係を調べるための予備実験を行い、各手法で検出可能なサイトが異なることを確認する。また、4 節における実験では SVM の判定信頼度に基づいて、明らかに違法・有害なサイトのみを検出した場合の判定精度と、判定があいまいであるサイトについてキーワードベース方式と組み合わせて判定を行った場合の判定精度を評価する。

3.2. HTML 部の抽出と文字列分割

Web サイトから HTML 要素を抽出、文字列に分割する方法について説明する。ここで HTML 要素とは HTML ファイルから本文テキストを除いた<>などで囲まれた部分とする。HTML ソースから本文テキストを抽出する方法については文献[8]や文献[9]などで提案されており、本稿では事前学習を必要とせず、計算量が少ないことを特徴とする文献[8]の手法を用いて、本文テキストと判定される部分を取り除いた HTML 要素を学習および判定に利用する。

次に、抽出した HTML 要素を文字列単位に分割する。区切り文字として、*\t , . / ! ” = % & { } [ ] _* などを設定し、HTML 要素を分割する。表 1 に HTML ソースと抽出した本文テキスト、本文テキストを除いた HTML 要素と HTML 要素を分割して抽出した文字列の例を示す。例えば、<a href> タグからは a, href, http, www, ipsj(サーバ名), or, jp, 10jigyo (フォルダ名やファイル名), html などが文字列とし

表 2　E(s)値算出に用いる文字列 s の出現回数

	文字列 s が出現	文字列 s が非出現	合計
有害サイト	$N_{11}(s)$	$N_{12}(s)$	N_p
無害サイト	$N_{21}(s)$	$N_{22}(s)$	N_n
合計	$N(s)$	$N(\neg s)$	N

表 3　文字列の出現回数と E(s)値の例

文字列	$N_{11}(s)$	$N_{12}(s)$	$N_{21}(s)$	$N_{22}(s)$	E(s)
S_1	100	1000	50	9850	122.9
S_2	10	1090	900	9000	-55.6
S_3	100	1000	900	9000	-2.0

て抽出される。

3.3. 違法･有害な文字列の抽出

学習用 Web サイトにおいて違法・有害サイトの HTML 部に偏って出現する文字列を自動的に抽出する。抽出手法として文献[2]と同様の手法を用いる。文献[2]では、ある文字列 s が違法・有害なサイトに偏って出現する度合いを表す指標 E(s)を AIC(赤池情報量基準)[10]を用いて算出する。表 2 のように、ある文字列 s が出現する違法・有害サイト数 N_{11} と無害サイト数 N_{21}、文字列 s が出現しない違法・有害サイト数 N_{12} と無害サイト数 N_{22} の 4 つの値を学習用サイトに出現する全ての文字列について求める。文献[2]では文字列 s が違法・有害な文書に偏って出現する度合い E(s)を文献[11]の知見を元に、AIC の独立モデルに対する値 AIC_IM および従属モデルに対する値 AIC_DM を用いて、次のように定義している。

$$
\begin{aligned}
&N_{11}(s) / N(s) > N_{12}(s) / N(\neg s) \text{ のとき、} \\
&\quad E(s) = AIC\_IM(s) - AIC\_DM(s) \\
&N_{11}(s) / N(s) \leqq N_{12}(s) / N(\neg s) \text{ のとき、} \\
&\quad E(s) = AIC\_DM(s) - AIC\_IM(s) \qquad (1)
\end{aligned}
$$

ここで、AIC_IM(s), AIC_DM(s)はそれぞれ文献[10]の定義に従って、次の式で与えられる。

$$
\begin{aligned}
&AIC\_IM(s) = -2 \times MLL\_IM + 2 \times 2 \\
&MLL\_IM = N_p(s) \log N_p(s) + N(s) \log N(s) \\
&\qquad + N_n(s) \log N_n(s) \\
&\qquad + N(\neg s) \log N(\neg s) - 2N \log N \\
&AIC\_DM(s) = -2 \times MLL\_DM + 2 \times 3 \\
&MLL\_DM = N_{11}(s) \log N_{11}(s) + N_{12}(s) \log N_{12}(s) \\
&\qquad + N_{21}(s) \log N_{21}(s) + N_{22}(s) \log N_{22}(s) \\
&\qquad - N \log N \qquad (2)
\end{aligned}
$$

具体例として、違法・有害サイトに偏って出現する文字列 S_1 と無害サイトに偏って出現する文字列 S_2、偏りなく出現する文字列 S_3 の例を表 3 に示す。S_1 は違法・有害サイトに偏って出現する文字列であるため、有害度合いを表す指標 E(s)が正の値をとり、S_2 は無害サイトに偏っているため E(s)は負の値を取る。S_3 は偏りなく出現するため、E(s)は 0 に近い値となる。(この例では、AIC の独立モデルを用いるか、従属モデルを用いるかの違いにより-2.0 の差が生じる。)

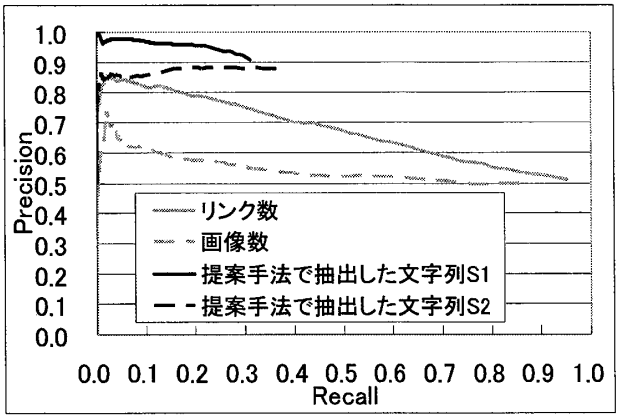

図 2　提案手法により HTML 部から抽出された文字列 S_1, S_2 と画像数、リンク数それぞれ N 回以上含むサイトを違法・有害と判定した際の性能比較

この手法により、違法・有害性の高いリンク先のサーバ名や違法・有害サイトで頻繁に用いられるポップアップなど Web ブラウザに特定の動作を要求する javascript 関数名などを自動的に抽出することができる。

抽出した違法・有害サイトの検出に役立つ各文字列について単独での性能を評価するための予備実験として、各文字列を N 回以上含む Web サイトを違法・有害と判定する方式において、N の値を変化させたときの再現率(Recall)と適合率(Precision)の関係を図 2 に示す。実験データとして人手により違法・有害または無害のラベルが付与された Web サイト各 1 万サイトを利用し、提案手法により統計的に抽出された違法・有害性の高い文字列 S_1,S_2 と文献[4]や文献[5]で有効とされている人手により観測された特徴量である画像数、リンク数の性能と比較する。

本稿では違法・有害サイト検出の再現率と適合率を判定対象となる Web サイト集合中の全違法・有害サイト数 All(本実験では 1 万サイト)、各手法で違法・有害と判定したサイト数 Judge、違法・有害と判定したうち、正しく違法・有害と判定できたサイト数 Correct を用いて、次のように定義する。

$$
\begin{aligned}
&Recall = Correct / All \qquad (3) \\
&Precision = Correct / Judge \qquad (4)
\end{aligned}
$$

図 2 から提案手法により得られた文字列 S_1,S_2 では同じ再現率においては、画像数やリンク数に比べて適合率が大きい傾向にあることが分かる。S_1,S_2 の再現率の最大値(1 回以上文字列が出現する違法・有害な Web サイトの割合)は画像数やリンク数に比べて低いが、複数の文字列を組み合わせることで向上することができる。このように適合率の高い特徴を持つ文字列を組み合わせることにより、提案手法では高精度を実現することが可能となる。一方、文献[4]や[5]で挙げられる人手による観測で有効性が高いとした画像数やリンク数などの特徴は適合率が低いため、組み合わせると全体の適合率が低下したり、利用する識別器のパラメータの最適化が複雑になり、未知データの識別に対する汎化性能が低下するなどの問題が生じる。

表 4　SVM の入力となる特徴量の例

	S_1	S_2	S_3	…	S_m	Label
サイト 1	N_{11}	N_{12}	N_{13}	…	N_{1m}	1
サイト 2	N_{21}	N_{22}	N_{23}	…	N_{2m}	0
…	…	…	…	…	…	…
サイト X	N_{X1}	N_{X2}	N_{X3}	…	N_{Xm}	0

3.4. SVM による学習と判定

3.3 節で抽出した違法・有害サイトの検出に役立つ文字列を組み合わせて SVM(Support Vector Machine)[12]を用いて違法・有害サイトの特徴を学習し、検出する。具体的には、抽出した文字列 $S_1, S_2, S_3, \ldots, S_m$ と各サイトにおける各文字列の出現回数 $N_1, N_2, N_3, \ldots, N_m$ からなる行列を SVM の入力として与える。学習フェーズでは加えて各サイトが違法・有害または無害を表すラベル Label も合わせて与えることで SVM を学習させる。表 4 に SVM の入力例を示す。

違法・有害サイトの検出に SVM を用いることの妥当性について述べる。本手法の利用シーンを考慮すると、学習データに対して正しい識別ができることよりも判定対象データ(未知のデータ)に対して汎化性能を示す識別器を利用することが望ましい。SVM は一般に汎化性能に優れていると言われており、本手法に適切と考えられる。予備実験として、提案手法の識別器として SVM と決定木を用いた識別器である C4.5[13]を用いた場合の性能を比較評価した。学習データとして人手により違法・有害または無害のラベルが付与された Web サイト 2 万サイト(違法・有害、無害各 1 万サイト)を用いて SVM と C4.5 をそれぞれ学習させ、判定対象となる学習用のサイトとは異なる 2 万サイト(違法・有害、無害各 1 万サイト)を判定し、F 値について評価した。SVM を用いた場合の F 値は 69.1%、C4.5 を用いた場合の F 値は 59.4%であり、SVM の方が本手法に適していることが期待される。C4.5 は著名な識別器であるが、この他に Neural Network[14]や Bayesian Filtering[15]なども有効性があると考えられ、これらを利用した際の性能の検証は今後の課題である。

また、SVM では判定の信頼度を計算することが可能であり、違法・有害または無害と判定する閾値をそれぞれ設定することが可能である。閾値を高く設定すれば再現率は低いが適合率は高くなる。閾値を低く設定すれば再現率は高くなるが、適合率は低くなる。4 節における実験では閾値を変化させたときの提案手法の再現率、適合率のトレードオフを評価する。

3.5. 提案手法とキーワードベース方式の特性

提案手法は Web サイトの HTML 部分のみを用いて判定を行うため、本文を対象として判定を行う既存のキーワードベース方式と組み合わせて利用することで、さらに高精度な判定を行うことが可能と考えられる。提案手法と従来手法[2]によって検出可能な違法・有害サイトの相関関係を調べるための予備実験を行った。

3.4 節における実験と同様に学習データとして人手により違法・有害または無害のラベルが付与された Web サイト 2 万サイト(違法・有害、無害各 1 万サイト)、判定対象データとして 2 万サイト(違法・有害、無害各 1 万サイト)を用いた。提案手法、従来手法それぞれにおいて、再現率が 10, 20, 30,…, 90(%) のとき、(1)提案手法でのみ違法・有害と判定したサイト数、(2)従来手法でのみ違法・有害と判定したサイト数、(3)両方の手法で違法・有害と判定したサイト数を図 3、図 4 に示す。再現率が大きくなるに従って、両方の手法で共通に違法・有害と判定したサイトの割合が増加するが、再現率 90%においても、各手法でのみ判定可能なサイトが存在することが分かる。この結果から提案手法と従来手法を組み合わせて利用することで、より多くの違法・有害サイトを検出することが可能と考えられる。4 節における実験では、提案手法において違法・有害と判定する SVM の信頼度閾値を高めに設定し、明らかに違法・有害なサイトを検出し、閾値に満たない判定があいまいであるようなサイトについてはキーワードベース方式を用いて判定するという手法の性能についても評価を実施する。

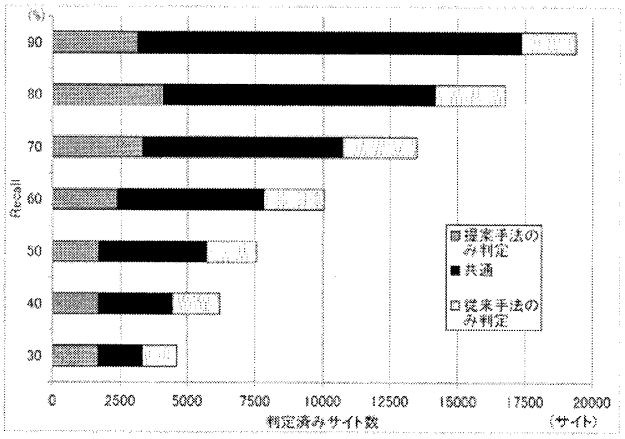

図 3　提案手法と従来手法において違法・有害と判定したサイト数

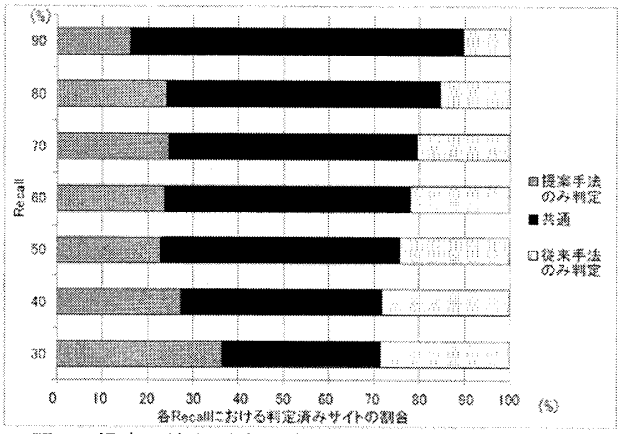

図 4　提案手法と従来手法において違法・有害と判定したサイト数(割合)

4. 性能評価実験

4.1. 実験の手順と環境

提案手法を実装し、キーワードベース方式の従来手法[2]との性能比較評価実験を実施した。実験環境と実験手順を下記に示す。

実験環境：計算機 1core 2.53GHz 64GB RAM Linux OS、提案手法で利用する SVM として Lib SVM[16]、従来手法で学習時に利用する形態素解析器として MeCab[17]を用いた。また提案手法、従来手法の実装には C 言語を用いた。

利用データ：Web サイト 4 万サイトを利用した。提案手法、従来手法それぞれ人手で違法・有害または無害のラベルを付与した学習用サイト 2 万サイト(違法・有害、無害各 1 万サイト)、判定対象サイト 2 万サイト(違法・有害、無害各 1 万サイト)を用いた。

評価指標：提案手法、従来手法において再現率と適合率を評価する。また、各手法において 1 サイトの判定に要する平均処理時間についても合わせて評価する。

実験手順：次に挙げる 5 つの手法の性能を比較評価する。(1)提案手法単独、(2)従来手法単独、(3)提案手法において判定の信頼度が閾値以上のサイトについては違法・有害と判定し、閾値以下の判定があいまいであるサイトについては従来手法を用いて判定する手法(以降では**複合手法**と呼ぶ)、(4)従来手法で抽出した違法・有害性の高い単語を提案手法と同様に組み合わせて SVM を用いて判定する手法を比較評価する。(1)の提案手法では、HTML から抽出した文字列 26 個を利用した。(2)の従来手法ではテキスト本文から抽出した単語 25000 個を利用した。(4)については提案手法と同量の 26 個の単語を利用した場合(4-a)と、10000 個の単語を利用した場合(4-b)についてそれぞれ評価した。

4.2. 実験結果

各手法における再現率と適合率の関係を図 5 に示す。(1)の提案手法と(2)の従来手法を比較すると、提案手法は 26 個という少数の文字列のみを利用したにも関わらず、再現率 50%以下の領域においては適合率が 90%以上と極めて高い適合率を実現している。再現率の高い領域においては従来手法の方が適合率は高くなる傾向が確認されるが、提案手法において有効性の高い文字列をさらに追加することで適合率、再現率の向上が期待される。

(3)の複合手法では再現率が 50%となるまで(1)の提案手法を用いて判定を行い、未判定のサイトを(2)の従来手法を用いて判定した。(1)の手法において性能が低下する再現率が高い領域においても性能が改善し、従来手法と比べて全ての再現率において高い適合率を実現することが分かった。特に再現率 70%においては従来手法と比べて適合率が 68.8%から 78.1%に 9.3%向上するなど、極めて効果的であることが分かった。F 値では(2)の従来手法が 70.6%であるのに対し、(3)の複合手法は 74.0%であった。

(4-a)の手法は(1)と同数の 26 個の単語をテキスト部から抽出したが、(1)よりも全体的に低い性能となった。これはテキスト部から特徴量として抽出した単語よりも HTML 部から抽出した文字列の方が、個々の特徴量の違法・有害サイトと無害サイトを識別する性能に長けているためと考えられる。(4-b)の 10000 単語を組み合わせて SVM を用いて判定する手法は再現率の高い領域において(1)の提案手法や(3)の複合手法よりも性能が高いことが分かった。

次に、判定に要した処理時間を表 5 に示す。(1)の提案手法と(2)の従来手法の処理時間はそれぞれ 3.85msec、3.57msec とほぼ同程度の処理時間となった。これは形態素解析のみを行った場合の処理時間と比べて半分程度であり、文献[3]のような高度な言語解析を行うキーワードベース方式と比べて高速であると言える。また、(4-b)の手法は再現率の高い領域において(1)の提案手法や(3)の複合手法よりも性能が高いが、多数の特徴量を組み合わせて判定を行うため、処理時間が大きくなる点が課題である。提案手法は少数の文字列でも比較的高精度を実現できるため、これらの問題を解決することができる点でも実用的である。

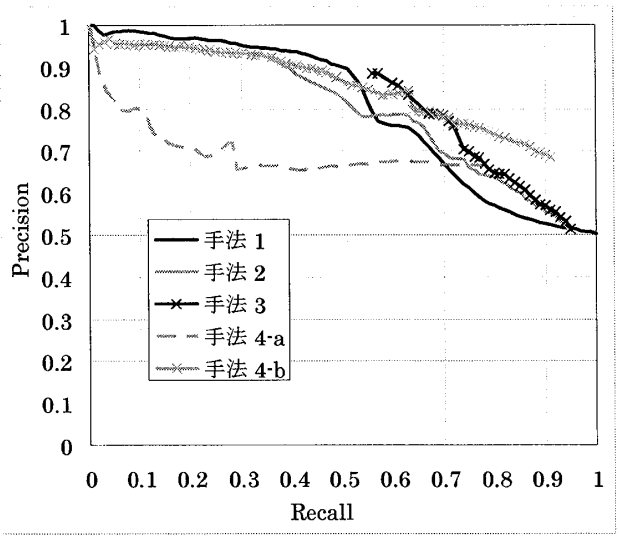

図 5　各手法における再現率、適合率の比較

表 5　判定に要した処理時間の比較

	1 サイトの判定に要した平均処理時間(msec)
手法 1(提案)	3.85
手法 2(従来)	3.57
手法 3(提案+従来の複合)	3.65
手法 4-a(従来 26 単語+SVM)	3.50
手法 4-b(従来 10000 単語+SVM)	12.12
形態素解析のみ(参考)	6.82

5. まとめ

本稿では高速かつ高精度に違法・有害サイトを検出するため、Web サイトの HTML を対象とした違法・有害サイト検出手法を提案した。提案手法では違法・有害サイトの HTML に偏って出現するような文字列を情報量基準に基づき統計的に抽出し、SVM を用いてこれらの特徴を組み合わせて違法・有害サイトの検出を行う。提案手法は Web サイトの本文の情報を利用しないため、既存のキーワードベース方式によって検出が困難なサイトも検出が可能であることを、各手法で違法・有害と判定するサイトの相関から検証した。性能評価実験においては、提案手法単体で利用した場合、再現率 50.0%、適合率 90.3%と極めて高い適合率が実現できることを確認し、さらに既存のキーワードベース方式と提案手法を組み合わせて判定を行う複合手法では、再現率 70.0%、適合率 78.1%を達成した。これは従来のキーワードベース方式の同程度の再現率における適合率と比較して 9.3%向上しており、極めて高性能なフィルタリングシステムを実現したといえる。

謝辞

本研究は、（独）情報通信研究機構の委託研究「高度通信・放送研究開発委託研究／インターネット上の違法・有害情報の検出技術の研究開発」の一環として実施した。

参考文献

[1] 総務省，“ブログの実態に関する調査研究”，2008, (URL:http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html#2008)

[2] 柳原正，松本一則，小野智弘，滝嶋康弘，“トピック判定における n-gram の組み合わせ手法の検討，”第 7 回. 情報科学技術フォーラム（FIT2008）論文集

[3] 井ノ上直己，帆足啓一郎，橘本和夫，“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. 84，no. 6，pp. 1158-1166，2001

[4] 本田崇智，山本雅人，川村秀憲，大内東，“Web サイトの自動分類に向けた特徴分析とキーワード抽出に関する研究，”情報処理学会研究報告 ICS，no. 78，pp.1-4，2005

[5] W. H. Ho and P. A. Watters, “Statistical and Structural Approaches to Filtering Internet Pornography”, in Proc. of IEEE International Conference on Systems, Man and Cybernetics, pp. 4792-4798, 2004

[6] 高橋功，三浦孝夫，“ハイパーリンクの共起性を用いたクラスタリング手法，”DEWS2005，1C-i12

[7] 唐門準，松尾豊，石塚満，“リンクに基づく分類のための ネットワーク構造を用いた属性生成，”情報処理学会論文誌，vol. 49，no. 6，pp. 2212-2223，2008

[8] 吉田光男, 山本幹雄, “教師情報を必要としないニュースページ群からのコンテンツ自動抽出”, 日本データベース学会論文誌, vol.8, no.1, pp.29-34, 2009,

[9] S. H. Lin and J. M. Ho, “Discovering Informative Content Blocks from Web Documents”, In Proc. of ACM SIGKDD, pp. 588-593, 2002

[10] 鈴木義一郎，情報量基準による統計解析入門，（株）講談社サイエンティフィク（編），pp.80-96，（株）講談社，東京，1995

[11] K. Matsumoto and K. Hashimoto, “Schema Design for Causal Law Mining from Incomplete Database,” Proc. of Discovery Science: Second International Conference(DS'99), pp. 92-102, 1999

[12] C. Cortes and V. Vapnik, “Support-Vector Networks, Machine Learning,” pp.273-297, 1995

[13] J. R. Quinlan, “C4.5: programs for machine learning, Morgan Kaufmann,” 1993

[14] S. Haykin, “Neural Networks: A Comprehensive Foundation,” Prentice Hall PTR, 1998

[15] D. J. Hand, H. Mannila and P. Smyth, “Principles of Data Mining,” The MIT Press, 2001

[16] R. Fan, P. Chen and C. Lin, “Working set selection using the second order information for training SVM,” Journal of Machine Learning Research, vol. 6 pp. 1889-1918, 2005.
(URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

[17] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to japanese morphological analysis,” Proc. of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004) pp. 230–237, 2004
(URL: http://mecab.sourceforge.net/)