5D-4

格要素の抽象化に基づく違法・有害文書検出手法の提案と評価

池田　和史†　　柳原　正†　　松本　一則†　　滝嶋　康弘†

† KDDI 研究所　〒356-8502　埼玉県ふじみ野市大原 2-1-15

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・中傷を含む書き込みなど、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。

違法・有害または無害と人手により判定された学習用文書における単語の出現頻度の偏りを用いて違法・有害判定のためのキーワードリストを自動生成する手法[1]も提案されているが、キーワードが文中でどのように利用されるかを考慮しないため、違法・有害情報の高精度な検出は困難である。例えば「爆破」という単語は「駅を爆破する」のような犯罪予告に用いられる一方、「炭鉱を爆破する」のように一般的な文書でも用いられる。

本稿では、文書から係り受け関係にある文節組を抽出し、違法・有害性との関連を学習すると共に概念辞書を用いて文節組を拡張することで高精度に違法・有害情報を検出する手法を提案する。提案手法は大規模 Web 文書群を用いた性能評価実験において、従来手法と比べて F 値で最大 3.9%違法・有害情報の検出精度を向上させることを確認した。

2. 関連研究

文献[1]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検出し、それらをキーワードとして、違法・有害情報検出を行う。しかし、文書を形態素に分割して扱う手法では、形態素同士の関係を考慮しないため、「爆破」や「薬物」のような前後の文脈に依存して違法・有害か無害かが分かれるような形態素を含む文書を正しく判定することが困難である。

一方、文書検索の分野では検索語および検索対象文書における文節の係り受け関係を考慮することで、高精度な文書検索が実現できることが報告されている[2]。違法・有害情報の検出においても、係り受け関係の利用が有用であることは大いに期待できる。

3. 提案手法

3.1. 提案手法の概要

従来手法[1]と提案手法における違法・有害情報検出の概要を図 1 に示す。判定対象となる文書には違法・有害な文書と無害な文書が混在している。従来手法は違法・有害性の高い順にランキングされたキーワードリストを自動生成し、保有しており、閾値以上の違法・有害性を持つキーワードを含む判定対象文書を全て違法・有害、それ以外の文書を全て無害と判定する。従来手法が違法・有害と判定した文書には無害と判定した文書に比べて多くの違法・有害文書が含まれるが、一部の文書は「炭鉱を爆破する」のように、「爆破」という表現を含

Detection of Illegal and Hazardous Information Using Dependency Relations and Keyword Abstraction

† Kazushi Ikeda, Tadashi Yanagihara, Kazunori Matsumoto, Yasuhiro Takishima, KDDI R&D Laboratories Inc.

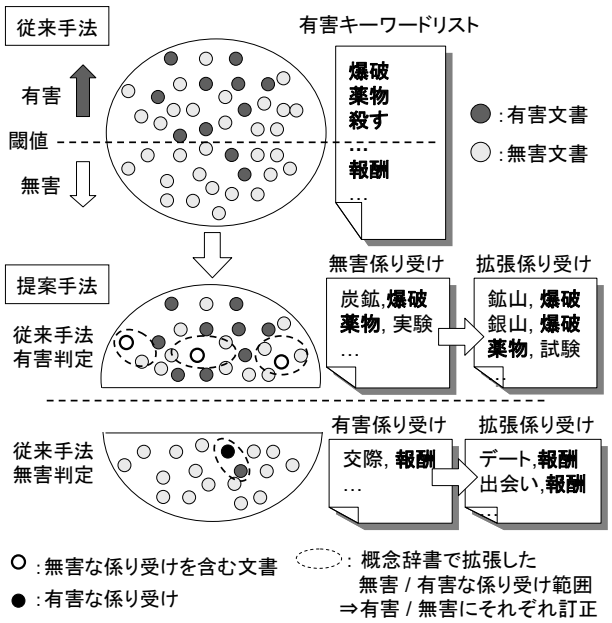

図 1　従来手法と提案手法の動作概要

んでいても無害である。提案手法では、従来手法で違法・有害または無害と判定された文書の中から、それぞれ無害または違法・有害性の高い係り受け文節組を検出し、従来手法の判定誤りを訂正することで精度を向上する。加えて、概念辞書を用いて係り受け文節組を抽象化し、拡張することでより多くの表現を検出する。

3.2. キーワードリスト生成手法

従来手法[1]では、学習用文書を形態素解析によって単語分割し、違法・有害な文書に偏って出現するような単語をキーワードリストに登録する。ある単語 w が違法・有害な文書に偏って出現する度合いを表す指標 E(w)は AIC（赤池情報量基準）[3]を用いて算出する。学習用文書に出現した任意の単語 w について、表 1 のように w が違法・有害または無害な文書に出現した回数 N_{11}, N_{21} および出現しなかった回数 N_{12}, N_{22} の 4 つの値を求める。文献[1]では E(w)を 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。

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

AIC_IM(w), AIC_DM(w)は文献[3]の定義から

$$AIC\_IM(w) = -2 \times MLL\_IM + 2 \times 2$$
$$MLL\_IM = N_{11}(w) \log N_{11}(w) + N_{12}(w) \log N_{12}(w) + N_{21}(w) \log N_{21}(w) + N_{22}(w) \log N_{22}(w) - N \log N$$
$$AIC\_DM(w) = -2 \times MLL\_DM + 2 \times 3$$
$$MLL\_DM = N(w) \log N(w) + N(\neg w) \log N(\neg w) + (N - N(w)) \log (N - N(w)) + (N - N_p) \log (N - N_p) - 2N \log N \qquad (2)$$

表 1　E(w)値算出に用いる単語 w の出現回数表

	単語 w が出現	単語 w が非出現	合計
有害文書	$N_{11}(w)$	$N_{12}(w)$	N_p
無害文書	$N_{21}(w)$	$N_{22}(w)$	N_n
合計	N(w)	N(¬w)	N

3.3. 係り受け文節組の抽出

学習用文書に対して、従来手法を用いて判定を行い、違法・有害と判定された文書からキーワードを含んでいる文に対して係り受け解析を行い、キーワードと係り受け関係にある文節組を全て取り出す（例えば図 1 で違法・有害なキーワード「爆破」と係り受け関係にある「炭鉱」の組を取り出す）。取り出した文節組 c に対し、表 1 と同様に違法・有害または無害な文書に出現した回数、出現しなかった回数をそれぞれ求める。このとき表 1 の N、すなわち文書数の総和は従来手法で違法・有害と判定された文書数となる。出現回数をもとに、3.2 節の(1)式および(2)式を用いて E(c)値を算出し、無害な文書に偏って出現する係り受け組をリストに登録する。同様に、従来手法において無害と判定された学習用文書を用いて、違法・有害な係り受け文節組を生成することもできる。この場合、従来手法の閾値以下のキーワード（図 1 では「報酬」）と係り受け関係にある文節組を求める。

3.4. 概念辞書を用いた拡張

3.3 節で取り出した文節組は少数の事例しか検出できないため、概念辞書を用いて拡張を行う。図 2 のように、文節組に含まれる単語をその単語の 1 つ上の概念以下に属する全ての単語と置き換えた文節組も同等の違法・有害性（3.3 節で求めた E(c)値）を持つとする。例えば図 1 において、閾値の設定により「爆破」が無害なキーワードとして扱われたとすると、提案手法によって「学校」と「爆破」の組は違法・有害性が高い係り受け文節組であると判定される。このとき、従来手法のキーワード「爆破」と組になる「学校」を抽象化する。これは「学校」の上位概念である「公共施設」の下位概念全て（「小学校」、「地下鉄」、「病院」など）を「学校」と置き換えても「爆破」と係り受け関係にある場合の違法・有害性は同程度になるという予測に基づいている。

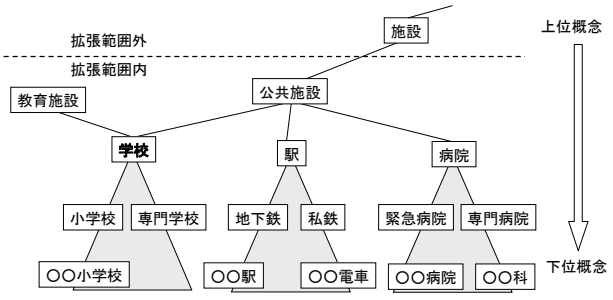

図 2　文節の抽象化手法

4. 性能評価実験

4.1. 実験の手順と環境

従来手法との性能比較評価実験の手順と環境を示す。

実験環境：計算機 1core 2.53GHz 64GB RAM Linux OS、形態素解析器として MeCab、係り受け解析器として Cabocha、概念辞書として EDR 電子化辞書を用いた。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：商用のブログ文書 80 万記事を利用した。提案手法、従来手法それぞれ学習用文書 40 万記事（違法・有害 4 万記事、無害 36 万記事）、評価対象文書 40 万記事(違法・有害 4 万記事、無害 36 万記事)を用いた。

評価指標：提案手法、従来手法において、Recall（再現率）、Precision（適合率）および F 値を評価する。

実験手順：

1. 従来手法において、違法・有害キーワードリストの閾値をいくつか設定し、Recall, Precision, F 値を評価する。
2. 1. に提案手法を適用し、従来手法の判定誤りを訂正し、Recall, Precision, F 値を評価する。

4.2. 実験結果

図 3 に実験結果を示す。提案手法では Recall は従来手法に比べ最大で 4.2%, Precision は最大で 2.0%, F 値は最大で 3.9%向上した。Recall の向上は従来手法で無害と判定された文書から違法・有害な係り受け文節組を検出し、正しい判定に訂正したためと考えられる。Precision の向上は従来手法で違法・有害と判定された文書から無害な係り受け文節組を検出し、正しい判定に訂正したためと考えられる。また提案手法では、学習文書中から得られた少数の係り受け文節組をもとに、概念辞書を用いて拡張したことで、より多くの表現を正しく判定することが可能になったと考えられる。

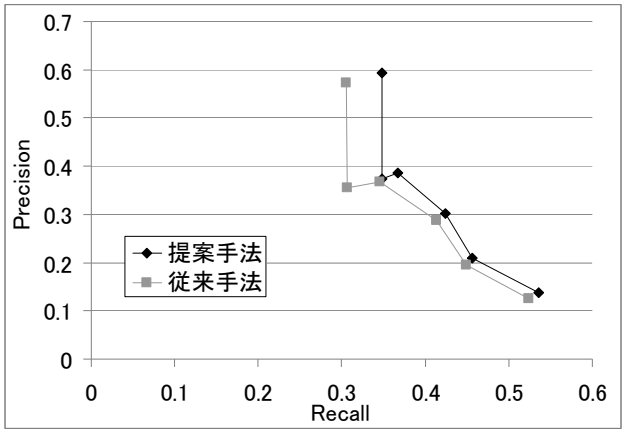

図 3　提案手法と従来手法の性能比較

5. まとめ

本稿では、文書から係り受け関係にある文節組を抽出し、違法・有害性との関連を学習し、さらに概念辞書を用いて文節組を拡張することで高精度に違法・有害情報を検出する手法を提案した。大規模 Web 文書群を用いた性能評価実験により、提案手法では F 値が 3.9%向上するなど、違法・有害判定の精度を従来手法に比べ性能を向上させることが分かった。

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] 柳原　他，“トピック判定における n-gram の組み合わせ手法の検討，” FIT2008，論文集

[2] 立石　他，“係り受け情報を利用した Web 上の日本語テキスト検索システム，”情報処理学会研究報告，vol. 98，no. 59，pp.47-54，1998

[3] 鈴木義一郎，情報量基準による統計解析入門，pp.80-96，（株）講談社，東京，1995