係り受け関係に基づく違法・有害情報の高精度検出方式の提案

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

E-mail: †{kz-ikeda, td-yanagihara, matsu, takisima}@kddilabs.jp

あらまし キーワードを用いて Web 上の違法・有害サイトを検出するフィルタリングシステムが普及しつつあるが、その多くは人手によりキーワードを設定している。機械的にキーワードを学習する手法も提案されているが、文中におけるキーワードの使われ方を考慮しないため、検出精度の向上が困難であった。本稿では、文書から係り受け関係にある文節の組を抽出し、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抽象化し、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。実際の Web から取得した違法・有害サイトを含む大規模 Web 文書群を用いて提案手法の性能評価を実施し、提案手法は従来手法に比べ F 値を最大 6.6%向上させることを確認した。

キーワード　情報フィルタリング、係り受け解析、概念辞書、キーワード検索

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. 2–1–15 Ohara Fujimino, Saitama, 356–8502 JAPAN

E-mail: †{kz-ikeda, td-yanagihara, matsu, takisima} @kddilabs.jp

Abstract　Keyword-based filtering systems for detecting illegal and hazardous information on Web site are spreading. Typical machine learning approaches for selecting those keywords ignore the ways of keyword usages in a sentence. In this paper, we propose a technique for automatically obtaining dependently related words that are biased and appear in illegal and hazardous documents. In addition, we also propose a technique to abstract and increase the effective keywords with thesaurus. Experimental results with large scale Web documents show that our method increases F value by 6.6% compared to the conventional method.

Keyword　Information Filtering, Dependency Relation, Thesaurus, Keyword Retrieval

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・有害な文書を自動的に検出するために、文書に特定のキーワードが含まれるか否かによって、文書が違法・有害であるかを判定するような情報フィルタリングシステムが普及しつつあるが、その多くは人手により違法・有害なキーワードを設定しており、拡張性に乏しい。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を用いて自動的にキーワードリストを生成する手法[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ているかを考慮しないため、違法・有害情報を高精度に検出することが困難である。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単語である一方、「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、既存のキーワードによる違法・有害情報検出手法において、違法・有害な文書を誤って無害と判定してしまう場合や、反対に無害な文書を違法・有害と判定してしまう場合について、キーワードを含む文節と係り受け関係にある文節の組を取り出し、違法・有害性との関連を学習することで、判定の誤りを減少させ、検出精度を向上する手法を提案する。加えて、概念辞書を用いてキーワードを含む文節と係り受け関係にある文節を抽象化することで、より多くの判定誤りを検出する手法を提案する。

実際の Web から収集した大規模 Web 文書群を用いて提案手法の性能を評価し、提案手法は従来のキーワードリスト自動構築手法と比べ、F 値で最大 6.6%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

Web サイトに含まれる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[1],[2]。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・有害判定を行う。文献[2]の手法では、学習用文書と評価対象文書の特徴ベクトルをそれぞれ求め、評価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、評価対象文書の違法・有害度合いを算出する。しかし、これらの手法では、文書を形態素に分解して扱っており、形態素同士の関係を考慮していない。そのため、「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[3],[4]。これらの手法では検索対象文書をあらかじめ係り受け解析しておき、ユーザから入力された自然語の検索文から取り出した、係り受け関係にある単語組を用いて文書検索を行う。これらの手法は、本稿とは目的が異なるが、高精度に違法・有害な情報を検出する上で、係り受け関係を利用することが有用であることは大いに期待できる。

一方、概念辞書を利用したクエリ拡張手法は古くから研究されており、様々な手法が提案されている[5],[6]。文献[5]では、クエリを抽象化する際に、ブーリアン式を組み合わせることで、適切な抽象度合いによる検索を実現するための手法を提案している。また、文献[6]では拡張される単語に多義性がある場合でも正しく抽象化を行うための手法を提案している。本稿では概念辞書を用いて文節組を抽象化する際、簡潔な手法を用いたが、これらの文献の知見を本稿で提案する手法に応用することで、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

従来手法[1]と提案手法における違法・有害情報検出処理の概要を図 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一様に分布している。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、閾値以上のキーワードを含まない文書を全て無害と判定する。例えば図 1 では閾値以上の違法・有害性を持つ「爆破」、「薬物」、「殺す」

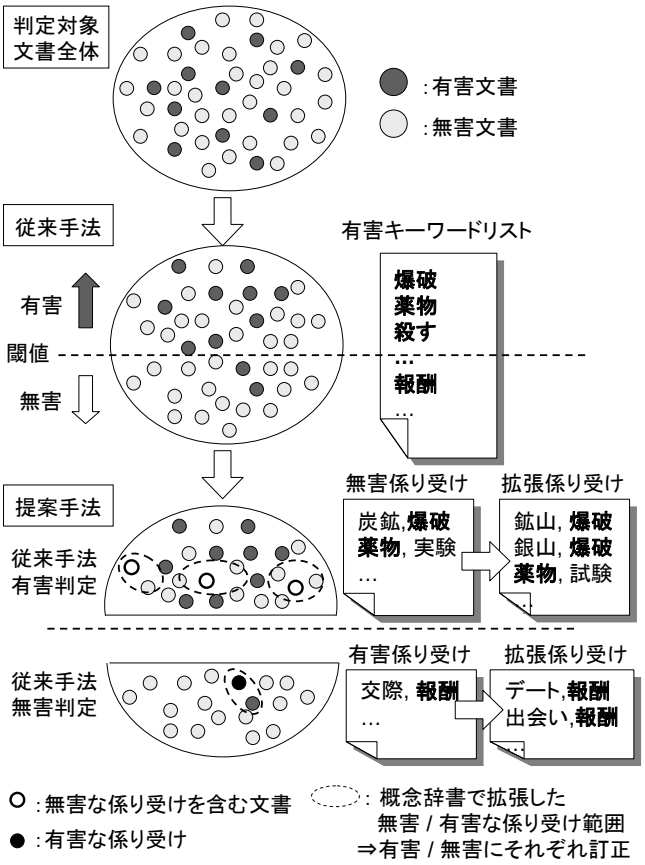


図 1　従来手法と提案手法の動作概要

といった表現を含む文書をすべて違法・有害としている。一方、閾値よりも違法・有害性の低い「報酬」などの表現はたとえ文書に含まれていても違法・有害と判定しない。

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」という表現を含んでいても無害であるし、「デートで報酬ゲット」のように「報酬」を含む違法・有害な文書も存在する。このように、違法・有害文書検出の再現率・適合率はトレードオフの関係にある。

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することで、従来手法における判定誤りを訂正し、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。

例えば、従来手法で「爆破」を含む文書は全て違法・有害と判定されたとき、提案手法では「炭鉱,爆破」という係り受け文節組である場合は無害であると判定を訂正する。加えて、概念辞書を用いて「炭鉱,爆破」を拡張した「鉱山,爆破」、「銀山,爆破」という係り受け文節組を含んでいる場合も同じく無害であると判定を

訂正する。同様に、従来手法において無害と判定された文書から違法・有害性の高い係り受け文節組を検出した場合も判定を訂正する。

以下では 3.2 節において従来手法における違法・有害キーワードリスト生成手法の概要を説明し、3.3 節において、提案手法における違法・有害または無害な係り受け文節組を生成する方法、3.4 節において、生成した係り受け文節組を概念辞書を用いて拡張する方法について述べる。

3.2. キーワードリスト生成手法

はじめに、従来手法[1]におけるキーワードリスト生成手法について述べる。文献[1]の手法では、人手により違法・有害または無害のラベルが付与された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）[7]を用いて算出する。表 1 のように、ある単語 w が出現する文書が違法・有害である場合の数 N_{11} と無害である場合の数 N_{21}、単語 w が出現しない文書が違法･有害である場合の数 N_{12} と無害である場合の数 N_{22} の 4 つの値を学習文書に出現した全ての単語について求める。文献[1]では単語 w が違法･有害な文書に偏って出現する度合い E(w)を文献[8]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。

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

ここで、AIC_IM(w), AIC_DM(w)はそれぞれ文献[7]の定義に従って、次の式で与えられる。

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

上記の計算により得られた違法・有害性の高いキーワードリストの一部を抜粋し、表 2 に示す。学習用の文書として Web サイト 22 万サイト（違法・有害 11 万サイト、無害 11 万サイト）に対して人手により違法・有害または無害のラベルを付与したものを利用した。ここでは、誹謗中傷や勧誘、成人向けの内容を含む文書を違法・有害と判定した。キーワードは違法・有害性の高さを表す E(w)値が高い順にランキングされているが、上位のランクであっても、無害文書が検出されるようなキーワードが含まれていることが分かる。このように、従来手法ではキーワードの前後の文脈を考慮しないため、違法・有害情報の高精度な検出が困難であった。

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

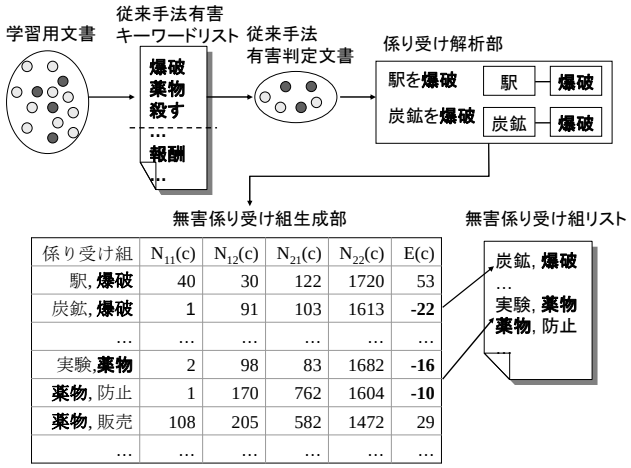
表 2　獲得した違法・有害キーワードリストの一部

Rank	キーワード	$N_{11}(w)$	$N_{12}(w)$	$N_{21}(w)$	$N_{22}(w)$	E(w)
…	…	…	…	…	…	…
10	女優	5802	102724	194	10833	6746
…	…	…	…	…	…	…
17	記事	1091	97615	3354	10517	4495
…	…	…	…	…	…	…
46	携帯	9253	99273	3259	10526	3167
…	…	…	…	…	…	…
106	スポンサー	2561	105965	708	10781	1129
…	…	…	…	…	…	…
110	アクセス	6573	101953	3361	10516	1105
…	…	…	…	…	…	…
163	アフィリエイト	1403	107123	292	10823	796
…	…	…	…	…	…	…

3.3. 係り受け文節組の抽出

従来手法において違法・有害または無害と判定された文書から、提案手法を用いて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するために、学習用の文書を用いて係り受け文節組リストを作成する方法について説明する。図 2 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学習する際の処理フローである。

学習用文書として、人手により違法･有害または無害のラベルが付与された文書を従来手法の違法・有害キーワードリストを用いて判定を行い、違法・有害と判定された文書から違法・有害なキーワードを含んでいる文に対してのみ係り受け解析を行い、違法・有害なキーワードを含む係り受け文節組を全て取り出す。取り出した係り受け文節組 c それぞれに対して、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



係り受け組	N_{11}(c)	N_{12}(c)	N_{21}(c)	N_{22}(c)	E(c)
駅, **爆破**	40	30	122	1720	53
炭鉱, **爆破**	1	91	103	1613	**-22**
…	…	…	…	…	…
実験, **薬物**	2	98	83	1682	**-16**
薬物, 防止	1	170	762	1604	**-10**
薬物, 販売	108	205	582	1472	29
…	…	…	…	…	…

図 2　係り受け文節組生成手法の概要

表 3　獲得した係り受け文節組と E(c)値

係り受け組	N11(c)	N12(c)	N21(c)	N22(c)	E(c)
…	…	…	…	…	…
女優,撮る	106	144651	2	72293	74.7
バラエティ, 女優	0	144757	3	72292	-4.52
プロデュース,女優	0	144757	2	72293	-2.31
…	…	…	…	…	…
スポンサー,出会い	14	144743	1	72294	20.3
提供, スポンサー	0	144757	6	72289	-1.83
スポンサー, 広告	0	144757	16	72279	-0.84
…	…	…	…	…	…
アクセス,偽る	7	144750	3	72292	16.1
アクセス, Copyright	0	144757	28	72267	-15.8
アクセス,ご案内	0	144757	27	72268	-9.20
…	…	…	…	…	…

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文節組をリストに登録する。

同様にして、従来手法において無害と判定された学習用文書から違法・有害な係り受け文節組を生成することができる。無害性の高い係り受け文節組の生成においては、従来手法の違法・有害キーワードリストの閾値以上のキーワード（図 2 では「爆破」、「殺す」、「薬物」）と係り受け関係にある文節組を求めたのに対し、違法・有害性の高い係り受け文節組では閾値以下のキーワード（図 2 では「報酬」）と係り受け関係にある文節組を求める。提案手法において得られた係り受け文節組の例を表 3 に示す。E(c)値の値が大きいほど、違法・有害性が高く、E(c)値の値が小さいほど無害性が強い。

表 4　抽象化で獲得した係り受け文節組と E(c)値

拡張前の係り受け組	拡張後の係り受け組	E(c)
…	…	…
女優,撮る	女優, 実写する	74.7
	女優, 流し撮りする	74.7
	女優, 裏撮り	74.7
…	…	…
スポンサー,出会い	スポンサー, 引合わす	20.3
	スポンサー, デート	20.3
	スポンサー, 交際する	20.3
…	…	…
アクセス, 偽る	アクセス, 偽造する	16.1
	アクセス, 模造する	16.1
	アクセス, 擬製	16.1
…	…	…

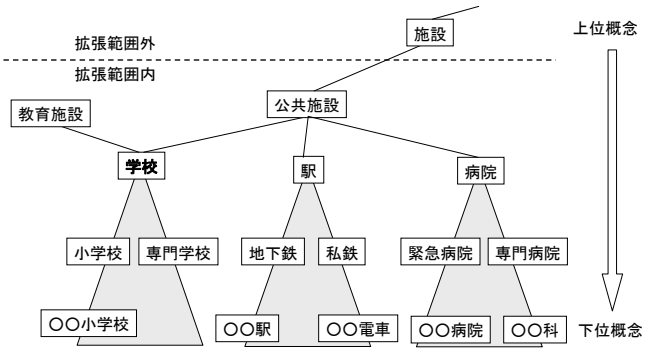


図 3　文節の抽象化手法

3.4. 概念辞書を用いた拡張

3.3 節で取り出した文節組について、概念辞書を用いて文節に含まれる単語を抽象化することでより多くの表現に適用可能とする。ここでは取り出した文節組のうち、違法・有害キーワードリストの単語を含まない方の文節を抽象化する。抽象化の手法としては図 3 に示すように、抽象化する文節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。例えば従来手法のキーワードリストに「爆破」が含まれており、閾値の設定により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、概念辞書を用いて従来手法のキーワードリストの単語「爆破」を含まない「学校」の方を抽象化する。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。ある係り受け文節組が複数の係り受け文節組から拡張によって得られた場合は、それらの出現回数の平均値を用い

て、E(c)値を算出する。例えば、「学校,爆破」と「駅,爆破」が得られていたとき、概念辞書を用いた拡張により「病院,爆破」が両方から得られたとすると、その違法・有害性を表す E(c)値は「学校,爆破」と「駅,爆破」の出現回数の平均から算出する。抽象化によって実際に得られた文節組の例を表 4 に示す。

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、従来手法との性能比較評価実験を実施した。実験の手順と実験環境を下記に示す。

実験環境：計算機 1core 2.53GHz 64GB RAM Linux OS、形態素解析器として MeCab[9]、係り受け解析器として Cabocha[10]、概念辞書として EDR 電子化辞書[11]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 24 万サイトを利用した。提案手法、従来手法それぞれ人手でラベルを付与した学習用文書 22 万サイト（違法・有害 11 万サイト、無害 11 万サイト)、評価対象文書 2 万サイト(違法・有害 1 万サイト、無害 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、Recall（再現率）と Precision（適合率）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変化させて違法・有害文書の検出を行い、従来手法の Recall, Precision のトレードオフについて評価する。
2. 1.のうち、いくつかの閾値の点を選択し、それぞれについて提案手法を用いて従来手法の判定誤りを訂正し、Recall, Precision を評価する。
3. 概念辞書を用いて 2.で作成した係り受け文節組を拡張したときの Recall, Precision を評価する。

4.2. 実験結果

はじめに、従来手法において違法・有害判定の閾値を変化させ、評価対象のブログ文書から違法・有害文書を検出した際の Recall, Precision を図 4 に示す。従来手法では違法・有害性の高い順にキーワードが整列されており、閾値が高いときは上位のキーワードのみが違法・有害判定に利用されるため、Recall は小さく、Precision は大きい。閾値を低く取ることで、利用されるキーワード数が増えるため Recall は大きくなるが、Precision は低下する。従来手法におけるいくつかの閾値について、提案手法を適用する。ここでは 8 つの閾値を選択した。各閾値において、利用されるキーワード数、Recall、Precision、F 値は表 5 のようになる。それぞれの場合について、提案手法を適用した。

図 5 に従来手法と提案手法である係り受け文節組のみを用いた手法、概念辞書を用いて係り受け文節組を

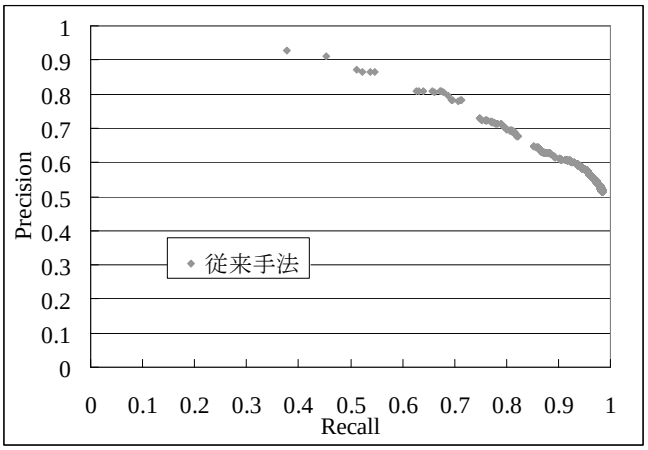


図 4　従来手法の Recall, Precision の関係

表 5　提案手法を適用する閾値の選択

	利用されるキーワード数	Recall	Precision	F 値
閾値 A	2	0.453	0.910	0.605
閾値 B	7	0.547	0.863	0.670
閾値 C	12	0.661	0.806	0.727
閾値 D	21	0.713	0.782	0.746
閾値 E	36	0.773	0.716	0.744
閾値 F	84	0.821	0.678	0.743
閾値 G	161	0.905	0.606	0.726
閾値 H	359	0.955	0.576	0.718

拡張した手法による違法・有害情報検出の Recall, Precision の関係を示す。提案手法では Recall, Precision 共に性能向上が見られた。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。

係り受け文節組のみを用いた手法では従来手法と比較して Recall は最大で 7.6%、Precision は最大で 2.0%向上した。概念辞書を用いて係り受け文節組を拡張した手法では Recall は最大 10.5%、Precision は最大で 3.2%向上した。これは提案手法において、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、新たに多くの表現を正しく判定することが可能になったためと考えられる。また、図 6 は F 値を比較したグラフであり、全ての閾値において、提案手法の方が高い値となっている。係り受け文節組のみを用いた手法では、最大で 4.8%の向上が見られ、概念辞書を用いた手法では最大で 6.6%向上した。

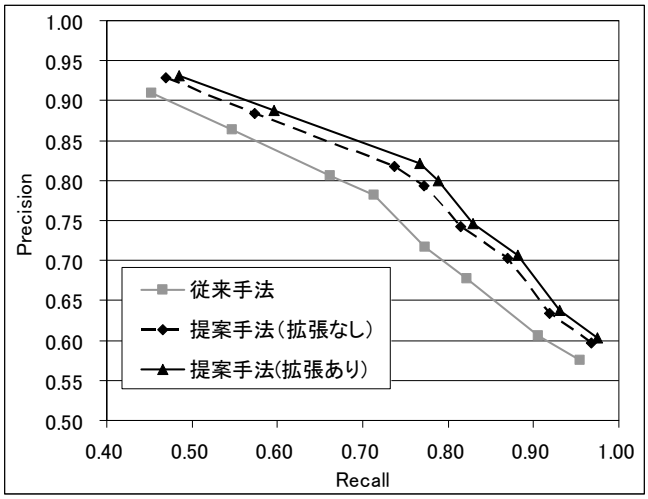


図 5　提案手法と従来手法の Recall, Precision の比較（破線は係り受け文節組のみを用いた手法。実線は概念辞書による拡張を用いた手法。）

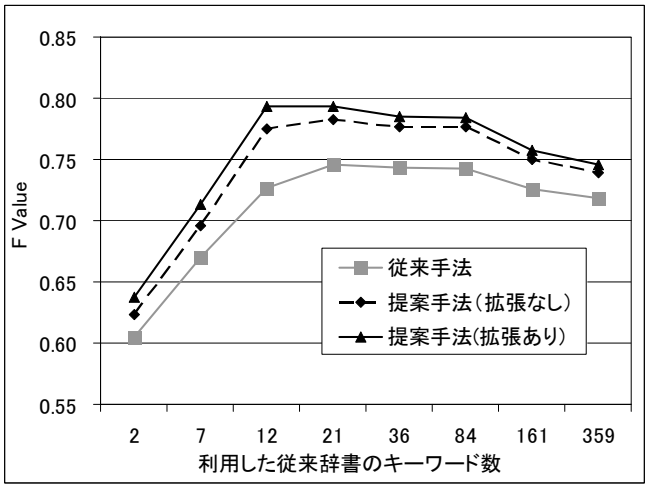


図 6　提案手法と従来手法の F 値の比較（破線は係り受け文節組のみを用いた手法。実線は概念辞書による拡張を用いた手法。）

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。実際の Web から取得した違法・有害サイトを含む大規模 Web 文書群を用いて提案手法の性能評価を実施した。

実験により、係り受け関係の抽出と概念辞書を用いた単語の抽象化を行うことにより、提案手法では F 値が 6.6%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

[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. 84，no. 6，pp. 1158-1166，2001

[3] 立石健二，大庭直行，峯恒憲，雨宮真人，“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

[4] 新美和彦，兵藤安昭，池田尚志，“係り受け関係を用いる高精度全文検索，”情報処理学会全国大会論文集，vol. 55，no. 3，pp.121-122，1997

[5] 吉岡真治，原口誠，“検索語の網羅性に注目した汎化概念により検索語選択支援を行う情報検索システムの研究，”人工知能学会論文誌，vol. 20，no. 4，pp. 270-280，2005

[6] Y. Liu, P. Scheuermann, X. Li, and X. Zhu, “Using WordNet to Disambiguate Word Senses for Text Classification,” Proc. of International Conference on Computational Science (ICCS 2007), Part III, LNCS 4489, pp. 780-788, 2007

[7] 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

[8] 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

[9] 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/)

[10] 工藤拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析，”情報処理学会論文誌，vol.43，no.6，pp.1834-1842，2002.
(URL: http://chasen.org/~taku/software/cabocha/)

[11] 独立行政法人情報通信研究機構，“EDR 電子化辞書，”(URL:
http://www2.nict.go.jp/r/r312/EDR/J_index.html)