O-030

教師データの投稿年代を考慮した有害情報の判定手法に関する検討
An Investigation for Detection Method of Harmful Information Considering Period of Posting for Supervised Data

岡 慎一郎 † 中村 健二 ‡ 小柳 滋 ‡
Shinichiro Oka Kenji Nakamura Shigeru Oyanagi

1. はじめに

インターネットには多種多様な情報が公開されており，その中には，未成年の出会いに関する情報や薬物取引情報などの有害なものも多く含まれている．これらの有害情報をフィルタリングする研究として，教師あり学習による分類手法を用いた取り組みが提案されている．例として，語の共起情報に基づく手法 [1]，係り受け関係に基づく手法 [2]，HTML 要素に基づく手法 [3] がある．一般的に，教師あり学習による分類手法では，有害情報の特徴を学習するための教師データが必要であり，判定精度は，教師データの量や質に依存している．特に，有害情報の特徴の学習を考えた場合には，新語や流行語への対策を念頭に入れる必要があり，投稿年代を考慮して教師データを準備することが重要であると考えられる．そこで，著者らは，投稿年代を考慮した有害情報の判定手法の開発を目指し，教師データの投稿年代の違いにより発生する判定精度の変化を検証する．また，インターネット上での出会いから，児童買春などの犯罪が増加していることから，本研究では，出会いに関する情報を有害情報，それ以外を無害情報と定義する．

2. 研究の内容

2.1 研究の全体像

投稿年代を考慮した有害情報の判定手法は，有害情報を判定する識別器を投稿年代毎に用意し，それらの識別器を組み合わせて一つの識別器を構築することで，判定対象の年代を問わず，高精度に判定する手法である．本研究は，そのための事前検討であり，教師データの投稿年代を考慮することで，判定精度の向上が可能かを検討する．

2.2 研究の概要

本研究では，教師データの投稿年代の違いにより発生する有害情報の判定精度の変化を明らかにするため，次の 2 つの内容について調査する．

- 調査 1：教師データの投稿年代と判定データの投稿年代に差がある場合，どの程度，判定精度に影響を与える可能性があるのか
- 調査 2：時代に応じて変化する有害語（以下，流行有害語）が判定データに含まれる場合，どの程度，判定精度に影響を与える可能性があるのか

これら 2 つの内容について明らかにすることにより，教師あり学習による分類手法を用いて有害情報を判定する際に，準備するべき教師データの特性を把握することができる．

†立命館大学大学院理工学研究科，Graduate School of Science and Engineering，Ritsumeikan University
‡立命館大学情報理工学部，College of Information Science and Engineering，Ritsumeikan University

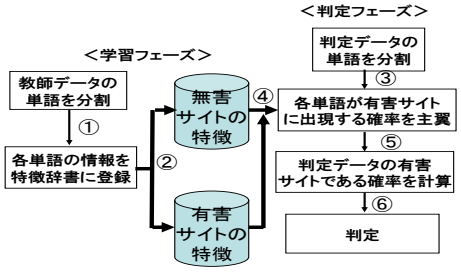

図 1: 処理の流れ

2.3 有害情報識別器の概要

本研究では，有害情報の分類手法として，Gary Robinson Fisher 方式 [4] を用いる．有害情報のフィルタリングシステムの概要を図 1 に示す．図に示す通り，有害情報の特徴を学習する学習フェーズと有害情報を判定する判定フェーズに分かれる．各フェーズの詳細を次に示す．

2.3.1 有害情報の特徴の学習

本処理では，まず，教師データに含まれる文書情報を形態素解析し，品詞が名詞である単語のみを抽出する．形態素解析器には MeCab を用いる．次に，任意の単語 $token_n$ が有害教師データに出現する確率 $p(token_n)$ を式 1 にて算出する．

$$p(token_n) = \frac{\frac{b_{token_n}}{SITE_{bad}}}{a\frac{g_{token_n}}{SITE_{good}} + \frac{b_{token_n}}{SITE_{bad}}} \quad (1)$$

式 1 において，g_{token_n} は任意の単語 $token_n$ が含まれる無害教師データの件数，b_{token_n} は任意の単語 $token_n$ が含まれる有害教師データの件数，$SITE_{good}$ は無害教師データの総数，$SITE_{bad}$ は有害教師データの総数を表す．

2.3.2 有害情報の判定

本処理では，判定データが有害であるか否かを判定する．まず，判定データに含まれる文書情報を形態素解析し，品詞が名詞である単語のみを抽出する．次に，各単語の有害確率 $f(token_n)$ を式 2 にて算出する．

$$f(token_n) = \frac{s \cdot x + n \cdot p(token_n)}{s + n} \quad (2)$$

式 2 において，x は任意の単語 $token_n$ の予測確率，s は予測確率 x に与える強さ，n は任意の単語 $token_n$ が出現する教師データの件数を表す．一般的に，$x = 0.5$，$s = 1$ が用いられる．次に，$f(token_n)$ の値を用いて，判定データが有害である確率 I を式 5 にて算出する．I の値が閾値よりも高ければ有害情報，低ければ無害情報と判定する．

$$H = C^{-1}(-2\ln \prod_{token_n} f(token_n), 2n) \quad (3)$$

$$S = C^{-1}(-2\ln \prod_{token_n} (1 - f(token_n)), 2n) \quad (4)$$

$$I = \frac{1 + H - S}{2} \quad (5)$$

3. 実証実験

3.1 実験内容

本実験では，教師データの投稿年代の考慮から発生する判定精度の変化を明らかにするために，第 2.2 節で示した 2 つの調査内容についての実験を行う．

1 つ目の調査内容を明らかにするために，年代毎の識別器を構築し，それらの識別器を用いて，各年代の投稿記事を判定した際の精度を評価する．具体的には，まず，投稿年代の異なる記事を 5 セット (2001，2003，2005，2007，2009 年の有害記事 400 件・無害記事 400 件) を準備する．そして，各年代それぞれ，600 件 (有害 300 件，無害 300 件) ずつ無作為に抽出し，年代毎の識別器を構築する．最後に，各年代の識別器を用いて，残りの 200 件 (有害 100 件，無害 100 件) を判定し，その精度を評価する．

2 つ目の調査内容を明らかにするために，無作為に収集した投稿記事から構築した識別器を用いて，流行有害語が含まれる投稿記事を判定した際の精度を評価する．具体的には，無作為に抽出した投稿記事 1,000 件 (有害 500 件，無害 500 件)(以下，通年教師データ）と流行有害語が含まれる投稿記事 1,000 件 (有害 500 件，無害 500 件)(以下，流行教師データ) を準備し，それぞれの識別器を構築する．そして，それぞれの識別器を用いて，無作為に抽出した投稿記事 500 件 (有害 250 件，無害 250 件)(以下，通年判定データ) と流行有害語が含まれる投稿記事 500 件（有害 250 件，無害 250 件)(以下，流行判定データ）を判定し，その精度を評価する．

3.2 実験結果と考察

表 1 に調査 1 の実験結果を示す．2001 年の記事の判定結果を確認すると，2001 年の記事の判定精度は，教師データと判定データの投稿年代の差が縮まるにつれて向上していることがわかる．これは，教師データと判定データに含まれる単語が類似しているためと考えられる．また，各年代の記事の判定精度が最良のものを確認すると，2007 年，2009 年の教師データを用いた識別器によるものであることがわかる．これは，2007 年，2009 年の教師データには，過去に利用された単語を多く含むためと考えられる．これらのことから，2007 年，2009 年の教師データを用いた識別器を構築することで，一定の精度を保つことができることがわかった．しかし，2001 年の記事の判定結果を確認すると，最良の判定結果は 2001 年の教師データを用いた識別器によるものであることがわかる．これは，年代差が大きいため，2007 年，2009 年の教師データに 2001 年の単語が含まれていないためであると考えられる．

表 2 に調査 2 の実験結果を示す．通年有害判定データの判定結果を確認すると，どちらの教師データを用いた場合も，高精度に判定できていることがわかる．これは，2 つの教師データに通年判定データの単語が多く含まれるためと考えられる．

次に，流行有害判定データの判定結果を確認すると，流行教師データを用いた識別器の方が，高精度に判定できていることがわかる．これは，通年教師データには，流行有害語のデータが含まれていないためであると考えられる．これらのことから，流行有害語に基づく記事を収集することで，高精度に判定可能な識別器を構築できると考えられる．

表 1: 調査 1 の実験結果（F 値）

		教師データ				
		2001	2003	2005	2007	2009
判定データ	2001	0.940	0.850	0.745	0.780	0.670
	2003	0.935	0.795	0.680	0.970	0.940
	2005	0.960	0.895	0.760	0.975	0.975
	2007	0.927	0.932	0.802	0.982	0.982
	2009	0.950	0.885	0.720	0.970	0.990

表 2: 調査 2 の実験結果

			教師データ					
			通年			流行		
			適合率	再現率	F 値	適合率	再現率	F 値
判定データ	通年	無害	0.968	0.852	0.906	0.951	0.868	0.907
		有害	0.871	0.976	0.920	0.897	0.976	0.934
	流行	無害	0.899	0.860	0.879	0.914	0.860	0.886
		有害	0.846	0.884	0.864	0.901	0.956	0.928

4. おわりに

本研究では，投稿年代を考慮した有害情報の判定手法のための事前検討として，教師データの投稿年代を考慮することで，判定精度の向上が可能かを検討した．

実証実験の結果より，教師データの投稿年代を考慮することで，判定精度の向上が可能であることが明らかとなった．今後は，本研究で得た結果をもとに，次の段階の投稿年代を考慮した有害情報の判定手法の研究に取り組む予定である．

参考文献

[1] 菊池琢弥，内海彰：語の共起情報に基づく有害サイトフィルタリング手法，第 9 回情報科学技術フォーラム講演論文集，No.2，pp.1-6 (2010).

[2] 池田和史，柳原正，松本一則，滝嶋康弘：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，DEIM Forum 2010，電子情報通信学会 (2010).

[3] 池田和史，柳原正，松本一則，滝嶋康弘：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，第 9 回情報科学技術フォーラム講演論文集，No.2，pp.7-12 (2010).

[4] Robinson，G.：A Statistical Approach to the Spam Problem，Lnux Journal，Vol.107，pp.58-64，Specialized Systems Consultants (2003).