K-029

「いじめ語」検出による学校裏サイト監視支援システム

Development of Monitoring System for School Unofficial Web-site with Detecting Bully Words

浅田 太郎 †　竹田 麻友子 ‡　吉冨康成 †　田伏 正佳 †
Taro Asada　Mayuko Takeda　Yasunari Yoshitomi　Masayoshi Tabuse

1. 緒言

近年、インターネットの普及により電子掲示板等、インターネット上での交流が盛んになった。それと共に「学校裏サイト」の数は 2008 年には全国で 3 万 8260 に達し、インターネット上で特定の人の誹謗中傷を行う「ネットいじめ」の件数も年々増えてきている[1]。また、近年「学校裏サイト」を監視する企業も現れているが、費用等の問題で企業に依頼することができず、また、多忙な先生や父兄が検索・監視を行うのも困難である。

そこで、本研究では、「ネットいじめ」監視の負担を軽減するため、形態素解析ツール MeCab[2]を利用して文章解析を行うことでインターネット上に書き込まれている「いじめ語」を検出するシステムを開発した。

2. 形態素解析ツールと「いじめ語」辞書登録

形態素解析とは文を形態素(言語で意味を持つ最小単位)に分割し, それぞれの品詞を判別することである。本研究では形態素解析ツールとして、MeCab を使用した。

学校裏サイトでいじめに繋がる「ウザイ」、「キモイ」、「死ね」等の誹謗中傷語を検出対象にする。書籍やインターネット上から実際使用されている誹謗中傷語を収集し、MeCab 辞書に品詞名を「いじめ語」として登録した。また、例えば「消えろ」に対して、「キエロ」、「きえろ」のような同意とみなせる単語の登録も行った。現在「いじめ語」228 語を登録している。「いじめ語」を辞書登録した MeCab による「いじめ語」検出例を図 1 に示す。

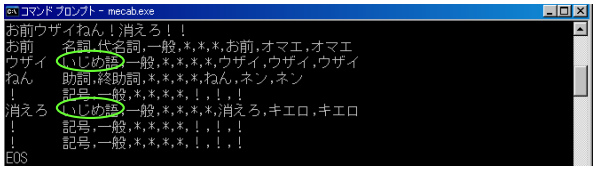

図 1　MeCab による「いじめ語」検出例

3. 処理概要

3.1　開発環境

システムは以下の開発環境で構築した。

- OS：Windows7
- 使用プログラム言語：Microsoft Visual C# 2008
- PC ： DELL　OPTIPLEX780(CPU ： Intel Core 2 Duo E8400 3 GHz , メモリ : 4 GB)

†京都府立大学 大学院 生命環境科学研究科，Graduate School of Life and Enviromental Sciences, Kyoto Prefectural University
‡シスメックス, Sysmex Corp.

3.2　処理の流れ

本システムの処理の流れを図 2 に示し、実装フォーム(注釈付)を図 3 に示す。

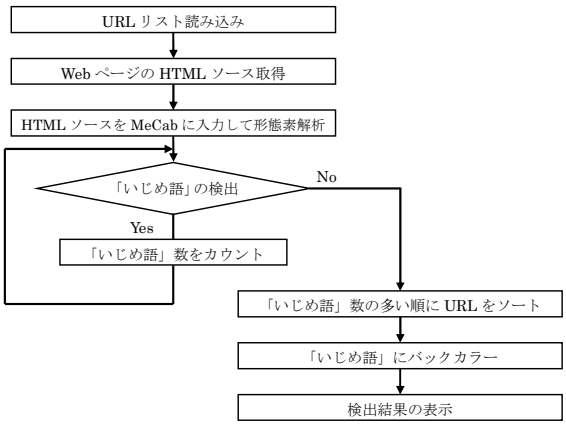

図 2　処理の流れ

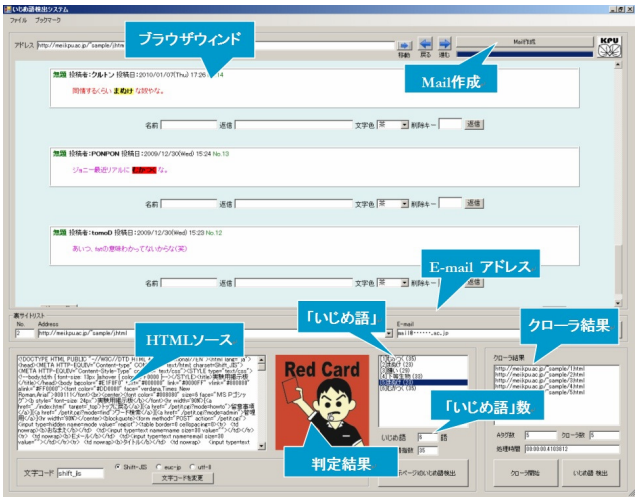

図 3　システムの実装フォーム

3.3　URL リスト

「いじめ語」検出対象のサイトの URL をアドレスバーに直接入力することもできるが、テキストファイルに「いじめ語」検出対象のサイトをリストアップし、URL リストとして、それを利用することもできようにした。

URL リストには、奇数行に学校裏サイトの URL、偶数行に直上の URL に該当する「学校や所轄の教育委員会」(以下、「関係先」と表記)のメールアドレスを記入する。メールアドレスは「いじめ語」検出の結果を関係先にメールで通報するときに使う。メールアドレスがわからな

い場合は、「no address」と記入する。メール送信機能については第 3 章 9 節で述べる。

3.4 HTML ソースの形態素解析

「いじめ語」検出を行うために、該当する Web ページの HTML ソースを取得する。取得した HTML ソースを MeCab に入力し、「いじめ語」検出を行う。このとき、“<TABLE>”、“</TABLE>”、“<TR>”、“</TR>”などの HTML でよく使用されるタグや、掲示板でよく使用される言葉（“投稿者”、“返信”など）を自動的に削除して MeCab に入力することで、形態素解析の処理時間を短縮している。削除対象としたタグ等の例を表 1 に示す。

表 1　削除対象としたタグ・言葉の例

<html>	<p>	<head>
</html>	Border	</head>
<title>	Width	text
</title>	Cellpadding	link
title	Cellspacing	vlink

形態素解析により検出された「いじめ語」をカウントし、該当の Web ページの「いじめ語」数として表示する。

3.5 「いじめ語」数による降順ソート

URL リストにリストアップされた学校裏サイトの中で、確認の必要性が高いサイトから順にチェックできるようにするため、「いじめ語」数の多い Web ページの順に URL をソートするよう設定した。

3.6 学校裏サイトの判定

2 章記載の「いじめ語」の辞書登録時、それぞれの「いじめ語」が与えるダメージの程度(以下、「いじめ語指数」と表記)も登録した。「いじめ語」検出を行った学校裏サイトのいじめ語指数を基に、学校裏サイトの通報等の対応必要性を 3 段階で判定する。

3.7 バックカラーの付加

検出された「いじめ語」に、いじめ語指数に応じて、バックカラーが付くように、HTML ソースを書き換え、ブラウザウィンドウに表示する（図 4）。この処理により、Web ページの「いじめ語」を容易に確認できる。

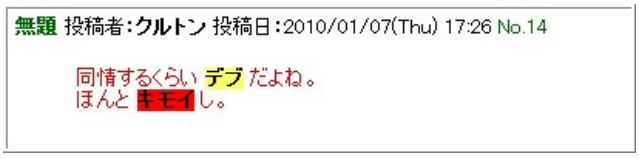

図 4　「いじめ語」にバックカラーのついた Web ページ

3.8 クローラ機能

既報[3]のクローラ機能を利用することで、ブラウザウィンドウ上に表示されている Web ページを起点として、リンクされている Web ページの「いじめ語」検出を可能にした。

3.9 メール送信

メール送信機能は、学校裏サイトから閾値以上のいじめ語指数の「いじめ語」が検出された場合などに、関係先にメールで通報する機能である。3 章 3 節記載の URL リスト登録時、メールアドレスを登録している関係先については自動的に宛先がメールに入力される。メールの本文には、「いじめ語」が検出された学校裏サイトの URL と、検出された「いじめ語」リストが自動的に記述される。

4. 「いじめ語」検出結果例

一般公開されている学校裏サイトリンク集「あげじゃぱん[4]」等のサイトから 100 スレッドを選び、本システムを使用して「いじめ語」検出を行った。検出された「いじめ語」とその数を表 2 に示す。

表 2　検出された「いじめ語」

「いじめ語」	検出数
死ね（しね、氏ね、死んで、タヒ 等）	48
うざい（ウザイ、うざ、ウザ）	28
カス（かす）	25
きもい（キモイ、きもすぎ、キモ 等）	22
馬鹿（バカ、ばか、バーカ、ばーか 等）	19
糞（クソ、くそ）	14
嫌い（キライ、嫌われ、嫌ってる）	14
消えろ	7
デブ	6
クズ	5
下等生物	3
キチガイ（基地外）	3
汚い	3
調子乗るな（調子に乗って）	2
ヲタ	1
殺す	1
潰し	1
ムカツク	1
無視	1
ざこ	1

5. 結言

学校裏サイトの HTML ソースの形態素解析を行い、「いじめ語」を検出するシステムを開発した。今後は、試行を通じて、実用化の課題を明らかにし、ユーザにより使いやすいシステムにしていく予定である。

参考文献

[1]文部科学省, “青少年が利用する学校非公式サイトに関する調査報告書”, 2008.
[2]京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト, http://mecab.sourceforge.net/
[3]礒部博行, 吉冨康成, “電子透かしを用いた著作権侵害防止システム”, 第 6 回情報科学技術フォーラム一般講演論文集, pp.51-52, 2007.
[4]あげじゃぱん, http://www.agejapan.com/