事例研究論文

Web スクレイピングを用いたネットパトロールの効率化

福地　彩[1], 鵜川義弘[2]
[1] 宮城教育大学大学院生活系専修, [2] 宮城教育大学情報処理センター

学校裏サイトの種類の中で、不特定多数が利用し書き込み頻度が高いインターネット掲示板は、大量の書き込みに問題発言が埋もれてしまう傾向があり早期に見つけ出して必要な対処を行うことが難しい。Web スクレイピング技術と、迷惑メール対策ソフトである POPFile を用いた分類技術を組み合わせて問題発言を抽出することで、いわゆるネットパトロール事業の一部に協力しているので、その方法を報告する。

キーワード : 学校裏サイト , ネットパトロール , Web スクレイピング , POPFile

1. はじめに

携帯電話の普及と、情報教育の遅れにより、児童生徒のネットワーク上でのトラブルが問題となっている。2008 年 4 月の文科省の調べでは学校裏サイト（学校の公式ホームページ以外で情報が書き込まれている電子掲示板など）が全国に 38,260 件あるとの調査 [1] がある。宮城県ネットパトロール事業をしている宮城県教育研修センターの話では、2010 年 1 月現在で、宮城県の学校裏サイトは、仙台市を除く 856 校について 14,069 件見つかっており、誹謗中傷、法規違反、プライバシーの流出などのトラブルが絶えないという。

特に、不特定多数が利用するいわゆる大型掲示板では、誹謗中傷をはじめとする問題発言は、1 日に 1 万件ほどもある大量の書き込みに埋もれてしまい、人手によるネットパトロールでは見つけ出すのが困難な状況にある。

そこで、Web スクレイピング技術を用いて問題発言を抽出し、それを前後の書き込みとあわせてメールとして送信し、迷惑メール対策ソフトである POPFile を用いて分類することで、問題発言のみを抽出した。結果をネットパトロール担当者に提供することで、その事業への支援とした。

2. Web スクレイピングによる書き込みの抽出

2.1 Web スクレイピングとは

Web スクレイピング（Web Scraping）とは、Web ページから個人が必要とする情報を自由に抽出できる技術のことを指す。本研究では、Perl を用いてシステムを構築したため、Web スクレイピングの方法として、Perl 用モジュール Web::Scraper[2] などのツールを用いて抽出する方法と、正規表現（文字列の特徴を表現する UNIX 表記）によりマッチした部分を抽出する方法とを検討したが、監視対象である掲示板の各スレッド（掲示板などで 1 つの話題に属する複数の書き込みをまとめたもの）の HTML ソースの構造が Web::Scraper での抽出に不向きであったため、正規表現により書き込みの抽出を行なっている。

2.2 監視対象

今回、監視対象とした掲示板は以下の 4 つである。

1. 宮城☆学生☆掲示板
2. 爆サイ .com - 東北版 -
3. ホストラブ（東北）
4. 新・石巻の高校掲示板

これらの掲示板は、宮城県教育研修センターで

これまで実際に目視で監視を行っていたものである。

2.3 書き込みの抽出

作成したプログラムは、以下の順序で書き込みの抽出を行なっている。

（i）スレッド一覧にアクセスし、スレッド番号を取得
（ii）取得したスレッド番号を元に各スレッドへ移動
（iii）各スレッドから書き込みデータを取得
（iv）過去の書き込みデータと比較し、差分を取り、新規書き込みを取得

掲示板では、通常、新しい書き込みが行われると、それが明確に分かるように、スレッド一覧ページ（図 1）の最上位にスレッドが表示される。そのため、スレッド一覧の TOP ページを見れば、新着書き込みがあったスレッドを把握することができる。

図 1 ホストラブ（東北）のスレッド一覧ページ

大抵の掲示板では、各スレッドにはスレッド番号を持っており、URL 内にその番号を記述することで、そのスレッドにアクセスできるようになっている。なお、前述した 4 つのサイトは全てスレッド番号を有していた。

各スレッドにアクセスすると、新しい書き込み内容が一覧として表示されるようになっている。(図 2)

このページのソースから、正規表現により、書き込み番号、書き込み日時、書き込み者名、書き込み内容を抽出し、データとして保存している。(図 3)

このデータは、スレッドの書き込みを全て抽出したものであるため、どれが新着書き込みかを判断することができない。そのため、新しくデータを保存したとき、直前に保存したものと比較を行い、差分を取ることで、直前のものに載っていないデータが新着書き込みであるとの判断を行っている。

図 2 スレッドの書き込み一覧

```
<NUM>846</NUM><DATE>2010/01/28 11:28</DATE><NAME>（匿名さん）</NAME>
<RES>退かぬ！媚びる！省みぬ！(^^)ｶｷｺしてる人消えてね(^^)</RES>

<NUM>847</NUM><DATE>2010/02/11 17:09</DATE><NAME>（匿名さん）</NAME>
<RES>退かぬ！媚びる！省みぬ！退かぬ！媚びる！省みぬ！退かぬ！媚びる！省みぬ！退かぬ！

<NUM>848</NUM><DATE>2010/02/11 17:10</DATE><NAME>（匿名さん）</NAME>
<RES>ワンサカワンサカ </RES>

<NUM>849</NUM><DATE>2010/02/11 17:10</DATE><NAME>（匿名さん）</NAME>
<RES>847(^^)しつこい</RES>

<NUM>850</NUM><DATE>2010/02/11 17:10</DATE><NAME>（匿名さん）</NAME>
<RES>レクサス</RES>

<NUM>851</NUM><DATE>2010/02/11 17:12</DATE><NAME>（匿名さん）</NAME>
<RES>848(^^)？ﾜﾝｻｶさんは誹謗中傷しないでしょ？勘違いｲｲとこかも(^^) </RES>
```

図 3 プログラムによる書き込みの抽出

書き込まれた内容が、問題あるものかそうでないかを判断するためには、データを書き込みごとに分割する必要がある。そこで、プログラムの最後の処理として、新着情報を書き込みごとに分割し、最終的にメールの形にして書き込み情報を後述する Gmail に転送している。

2.4 Gmail の利用

Gmail は Web 検索大手の Google が提供する無料の WebMail である。その検索機能は Web の検索で想像がつくとおり秀逸であり、かつ IMAP サーバとしての利用も可能である。筆者らは、掲示板に書き込まれた記事を Gmail に転送することで、掲示板記事の保存の受け皿とし、これを県の職員とメールフォルダーの検索を含めて共有することとした。ただ、1 日 1 万通を超えるメールの転送は Gmail でさえ迷惑メール送信サイトとして扱われたため、一旦学内のサーバに蓄積したのち、Gmail の「POP を使用したメッセージの確認」機能を用いることで Gmail 側から取得させるようにした。

3. POPFile による書き込みの分類

3.1. POPFile とは

前節で最終的に書き込みをメールとしたのは、POPFile[3] を用いて書き込みを分類するためである。POPFile は、ベイジアンフィルタを持つ自動メール分類ソフトウェアで、元々は迷惑メール対策として開発された。ベイジアンフィルタとはベイズ推定を用いたフィルタのことで、メールの分類を学習させることにより将来的に自動で分類が判定されるようになる。単語の発生頻度から分類先を判定しているため、学習量が増えるとその分だけ分類の精度も上がるという特徴を持つ。会員数 1500 万人を有する DeNA の Web サービス「モバゲータウン」でもこのベイジアンフィルタの技術を用いて書き込みのチェックを行っている。ベイジアンフィルタを持つメール分類ソフトウェアは複数存在するが、多くは迷惑メール対策としての利用に限定され、spam と ham とに分類することしかできない。しかし、本研究で採用した POPFile は、バケツと呼ばれる分類ジャンルを自由に複数設定することができるため、迷惑メール対策以外の用途にも利用可能である。

3.2 POPFile の設定

POPFile は通常では POP のプロキシとしての利用が想定されているが、ベータテスト版として IMAP での利用もできるようになっている。今回は、Gmail に蓄積されているメールを分類するため POP ではなく IMAP での設定を行った。設定は、「詳細設定」タブで、imap_enabled に値“1”を入力し、IMAP サーバに関する設定に Gmail アカウントに関する情報を入力する。　また、POPFile では「バケツ」と呼ばれる分類のジャンルを専用の Web インターフェースにて作成することができる。今回は、まだ実験段階ということもあり、分類のジャンルとして、監視を必要とする書き込みであることを表す「watch」、問題のない書き込みを表す「trash」を作成した。(図 4)

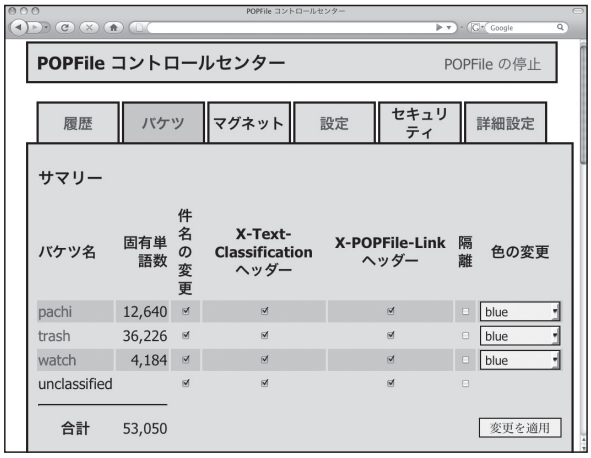

図 4 POPFile ジャンルの設定画面

3.3 POPFile による書き込みの分類

POPFile はコーパスと呼ばれる単語データを元に分類を判定するが、インストール直後はそのコーパスを持たないため、手動で分類を行うことで学習をさせ、コーパスのデータを充実させる必

要がある。学習は POP 経由の場合には、Web インターフェース（図 5）で行うが、IMAP では、Thunderbird など、通常のメールソフトを使って行う（図 6）。

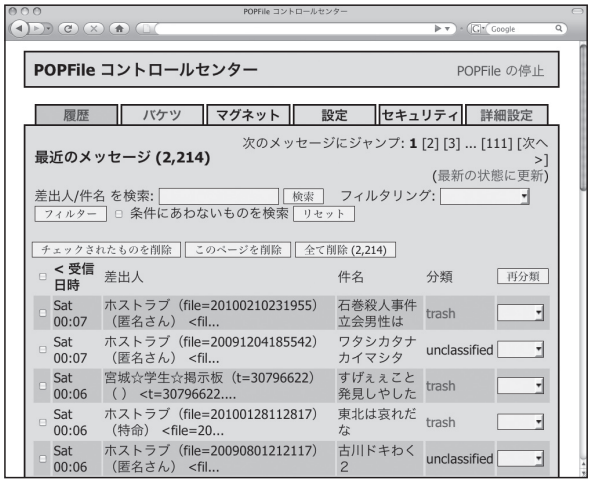

図 5 POPFile の学習（Web 画面）

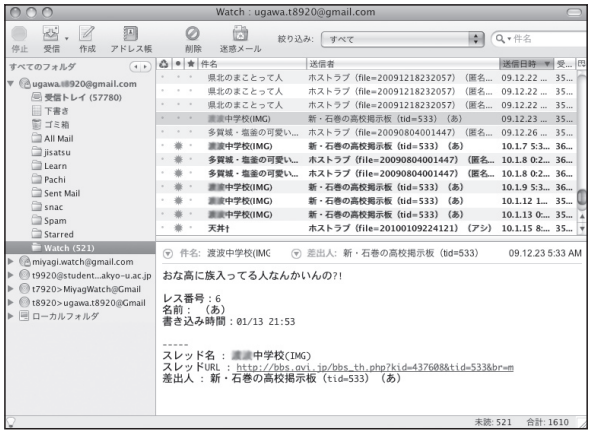

図 6 POPFile の学習（Tunderbird の場合）

具体的には未分類「unclassified」の場合、受信トレイに溜まっている状態なので、掲示板の書き込みの内容により「watch」または「trash」に移動することにより学習されコーパスが作られていく。徐々にそれをもとに分類の判定が行われ「watch」または「trash」に自動で分類されていくようになる。もし、その判定が誤っている場合は、正しい行き先フォルダーに手作業で分類することで再学習が行われる。分類精度が上がれば再学習の頻度は少なくなり、POPFile によって自動的に「watch」に分類されたメールだけをチェックすることで、注意すべき書き込みを監視することが可能になる。

実際に使用してみると、分類精度を上げるには、1 件の書き込みにある程度特徴的な単語が入っていることが必要であることがわかったため、Web スクレイピング時には、更新部分の 1 または数行だけでなく、掲示板のほぼ 1 ページに相当するスレッドごと収集するようにした。

4. 今後の課題

大量の書き込みの中から問題となる書き込みを見つけ出す方法は提供できたものの、最も重要で緊急を要する自殺、自傷系の発言については、これまで開発したシステムでは見付け出すことができない。それは、自殺、自傷の発言が、児童生徒のブログやリアル（ブログよりも頻繁に書き込まれる日記的記載）にあり、それらは、今回のシステムで行ったいわゆる定点観測による方法ではなく、前略プロフィールを起点とするリンクから派生する深いリンクをめぐる Web　Crowling 技術を使わなければ見つからないからである。児童生徒の命がかかる問題に対して情報と教育分野にたずさわる者として、早急に対応しなければならない大変重要な課題である。

[1] 文部科学省調査「青少年が利用する学校非公式サイト」http://www.mext.go.jp/b_menu/houdou/20/04/08041805/001.htm（2010 年 1 月 14 日アクセス）
[2] Web::Scraper http://search.cpan.org/~miyagawa/Web-Scraper-0.32/lib/Web/Scraper.pm（2010 年 1 月 14 日アクセス）
[3] POPFile http://getpopfile.org/docs/jp（2010 年 1 月 14 日アクセス）