学校非公式サイトの分析

文教大学大学院　情報学研究科　専任講師　池辺正典†
Masanori Ikebe

あらまし

インターネットの発達に伴い、誰もが容易に情報発信を可能にする社会が実現した。これにより、学校では、教育内容などの話題がインターネットに公開されるに至った。これらの Web サイトは学校非公式サイトと呼ばれ、児童・生徒・学生をトラブルに巻き込むという問題が発生する傾向にある。このため、教育機関から、学校非公式サイト対策への要望が高まっており、本研究は、対策準備段階として学校非公式サイトを分析したものである。

キーワード：学校教育、テキスト・マイニング

1. はじめに

学校非公式サイトとは、教育機関から教育を受ける側である児童・生徒・学生が主体となり学校についての情報発信を行う公式サイト以外の Web サイトを示す。学校非公式サイトは、多数の利用者が学校に関する情報を発信する形式が一般的であり、その媒体としては、BBS（掲示板）・Blog・SNS（Social Networking Service）のような利用者間のコミュニケーションを重視した CGM（Consumer Generated Media）に分類されるメディア形式が利用されることが多い。そして、文部科学省が 2008 年 3 月に公表した学校非公式サイトの調査[1]では、38,260 件の学校非公式サイトが発見されており、さらに他の利用動向調査によると、各学校は複数の学校非公式サイトを有するという結果が得られていることから、先の文部科学省の調査は、氷山の一角に過ぎないと考えられる。

このようにインターネット上に多数確認することができる学校非公式サイトの問題点としては、Web サイトの問題が現実世界へのトラブルに繋がるという点である。近年のニュースなどにおいても、学校非公式サイトでのいじめから児童・生徒が自殺に繋がるケースや大学においては、就職活動後の学生が企業による Web の調査において、問題行動を発見され内定取り消しに繋がるケースなど多様な問題が発生している。これらの問題は、いずれも個人が特定可能なことから発生する問題である。このため、当面の学校非公式サイトで発生する問題の対策として有効なものは、固有表現（特に氏名）についての対応である。そして、固有表現について教育機関が警戒することが重要であると考えられるが、この対策を実現している教育機関は非常に少ないと考えられる。

2. 固有表現の抽出

文書内から固有表現の抽出する手法としては、各文を名詞や動詞などの品詞単位に分割した上で、その接続関係から品詞特定を行う形態素解析が一般的な方法である。この手法により独自に収集を行った小中高の学校非公式サイト 1,000 件を分析した結果、15,968 件の個人名を取得することが可能であった。その内容を Web サイトの全単語からの比率を確認すると小学校では 1.43%、中学校では 1.63%、高等学校では 0.77%の単語が個人名であった。また、同様の 1,000 件の Web サイトの話題について、学校非公式サイトでよく話題に上る「友達募集」、「学校生活」、「誹謗中傷」、「日常会話」、「受験・テスト」、「悩み相談」、「恋愛関係」、「その他」の 8 つのカテゴリーへと目視により分類した結果、「誹謗中傷」の割合は、小学校で 15%、中学校で 40%、高等学校で 13%であり、先の個人情報とある程度の相関を確認することができる。このことからも固有表現への対策は、学校非公式サイトでの問題対策に繋がると考えられるが、形態素解析のみで学校非公式サイトの固有表現を抽出するには限界がある。

3. Web サイト解析時の問題

形態素解析のみでは、学校非公式サイトから固有表現を抽出することが困難である理由としては、辞書の問題と Web 表記の慣習の問題がある。

最初の問題として、形態素解析は、品詞の接続関係の判定などに辞書を用いるが、精度がよいとされる形態素解析用の辞書はいずれも新聞記事をベースに学習が行われているために、新聞記事のようなしっかりとした体裁の文書では非常に高い精度を得ることができるが、文の体裁が曖昧なものや統一性がないものに対しては精度が下がる傾向が

2009 年 6 月 31 日受付
† 〒253-8550 神奈川県茅ヶ崎市行谷 1100　m_ikebe@shonan.bunkyo.ac.jp
Graduate School of Information and Communication, Bunkyo University

ある。学校非公式サイトは、多数の利用者が情報発信を行う CGM 形式であるために、文の体裁が統一的でない。このため、解析用の専用の辞書作成し、さらに、文の体裁統一を行うための辞書を別途作成することで、精度向上に繋がると考えられる。

次の問題として、CGM などの利用者発信型の Web の慣習では、単語を正確に表記するのではなく、意図的に当て字やアルファベットや数字の略称などを利用する隠語の表現が多い。JC（女子中学生）・JK（女子高生）などは比較的有名な隠語表現であるが、このような隠語表現は、問題に繋がりやすい単語ほどよく用いられる傾向がある。これらの問題に対応するには、形態素解析の前に隠語表現が行われている単語についての類義語辞書を整備するなどの対策が必要であると考えられる。

4. Web サイトのカテゴリー分類

学校非公式サイトは、全てが問題ではなく、教育改善などに重要な手助けとなる情報発信が行われている Web サイトも少なくない。このため、学校非公式サイトへの問題対策では、主に「誹謗中傷」を目的とした Web サイトを抽出する必要がある。先の独自調査においても、「誹謗中傷」を目的とした学校非公式サイトが全体で 22.7%存在したが、このような問題に繋がることの多い Web サイトは、頻繁に移動を行うなどで一般からは発見を避けるような運営が行われ、人為作業による調査は非効率的であると考えられる。このため、プログラムによる自動的な判定により問題のある学校非公式サイトを抽出することが望まれるが、このようなカテゴリー分類手法は、ベクトル空間法[2]を用いる手法や SVM（Support Vector Machine）などを用いる手法[3]が多い.

ベクトル空間法とは、解析文書から単語の出現頻度と単語の特徴値を数値化し、仮想の多次元空間上にベクトルとして展開することで、その距離から類似度計算を行うものである。

そして、SVM を用いる手法とは、パターン認識の 1 手法であり 2 つの個体群の分離平面を求める SVM を応用することで、多クラスの分類をカテゴリー分類に利用するというものである。線形分離が不可能な固体群に対しては、他の特徴空間に写像することで対応する。

ここでは、ベクトル空間法を用いた学校非公式サイトの分類[4]を行った。その内容は、20 件の学校非公式サイトについて、小学校・中学校・高等学校・一般の Web サイトの 4 種類のカテゴリーに自動分類を行うというものである。その結果、80%の精度で分類が可能であった。今回のカテゴリー分類の前提条件として用いた学習データは、4 種類のカテゴリーが各 50 件の合計 200 件であり、単語の特徴値の上位 1,000 件・3,000 件・5,000 件でそれぞれ試行を行った。カテゴリー分類時には、単語の特徴上位の何件まで取得するかにより、精度に違いが見られるが、今回の分類では、それぞれの件数の試行において、結果が異なったのは 5%のみであり、今回の分類においては 1,000 件で充分であると考えられる。

しかし、同様の手法を用いて、「誹謗中傷」が行われている学校非公式サイトをカテゴリーとして抽出する場合には、特徴の違いが小さくなる可能性もあることから、さらなる精度向上が必要であると考えられる。具体的に追加する内容として想定される手法は、先に形態素解析の前処理として挙げた文書体裁の統一方法を充実させることや、類義語（隠語表現）に対応するために、特徴値を仮想空間に展開する際に LSI（Latent Semantics Indexing）などを利用して、類義語の特性を維持したまま、複数の単語を 1 平面に射影することで、隠語表現の精度向上に繋げるなどの発展が考えられる。

5. おわりに

インターネットというメディア媒体は、誰もが容易に情報発信できるという特性から急速な発展を遂げた。学校非公式サイトで現在注目を集めている内容は、今回述べたような問題の対策方法であり、これはインターネットの特性から発生している問題でもあると考えられる。そして、インターネットで近年注目されている研究課題としては、利用者が発信した情報を評価情報として抽出し、サービスの改善に繋げ、利用者の利便性を向上させるというものである。このため、現状の教育現場での問題対策以外の学校非公式サイトの活用方法として、教育機関の評価情報を抽出し、教育内容の改善に繋げるなどの活用方法を検討することも可能であると考えられる。教育機関として、学校非公式サイトに対応する場合には、問題対策という側面だけでなく、このような教育改善などの建設的な活用方法についても検討が必要であると考える。

［文 献］

1) 文部科学省：青少年が利用する学校非公式サイト等に関する調査について, http://www.mext.go.jp/b_menu/houdou/20/04/08041805/001.htm .
2) 池辺正典，田中成典，古田均，中村健二，小林建太: Web リンク構造解析と自然言語処理による組織関係の抽出についての研究，情報処理学会論文誌，情報処理学会, Vol.47, No.6, pp.1687-1695, (2006).
3) 矢田裕之，上原邦昭：Support Vector Machine によるテキスト分類における属性選択，情報処理学会論文誌, Vol.411，No.4，pp.1113-1123 (2000).
4) 池辺正典，佐久間拓也，川合康央，柳生和男，松本浩之：学校非公式サイトを活用した学校評価支援に関する提案，情報教育シンポジウム論文集，情報処理学会, Vol.2008, No.6, pp.193-200, (2008).

いけべ　まさのり
池辺　正典
1977 年生。関西大学大学院総合情報学研究科博士後期過程修了。2007 年 4 月より文教大学情報学部に着任。情報システム，データマイニングなどが専門。本情報学研究科では「ウェブコンテンツ演習」を担当。