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

池辺正典† 佐久間拓也† 川合康央† 柳生和男† 松本浩之†

近年の Web の普及は，容易に利用者を情報発信の枠組みに参加させることとなり，結果として，Web 上に特定のテーマを話題とする多人数が参加するコミュニティを生成するに至った．これらのコミュニティには，学校教育をテーマとする学校非公式サイトも含まれており，生徒や学生が日常的に学校教育に関する情報を提供しており，その影響力は強くなっている．そして，学校教育の改善や円滑な運営を行うためには，こうしたコミュニティから情報を取得する必要があるが，コミュニティの非公式という特性上，該当するコミュニティを発見するのは困難である．
そこで，本論文では，学校非公式サイトについての情報を自動的に取得することで，教育者に対して，教育関連の評価情報を継続的に提供することを目的とする．教育改善においては，前述のコミュニティ内での問題となる話題の発見，解決策の検討，問題解決への取り組みが行われる必要があるが，本論文では，最初に必要となる情報源の把握と継続的な監視に関する提案である．

Proposal for Supporting School Evaluation Using School Unofficial Website

MASANORI IKEBE† TAKUYA SAKUMA† YASUO KAWAI†
KAZUO YAGYU† HIROYUKI MATSUMOTO†

In recent years, the spread of Web facilitated the information sending by the user. As a result, the community that treated a specific theme was generated on Web. These contain the school unofficial website and students offer information that has the influence power about the education daily. It is necessary to acquire information from the community for an educational improvement and smooth management. The discovery of the community is difficult because the community is informal.
In this paper, information on the school unofficial website is automatically acquired. And the teacher obtains a continuous dissemination. An educational improvement is achieved by "Problem discovery", "Examination of the solution", and "Approach of the problem solving". This paper proposes a grasp of the source and a continuous watch.

1. はじめに

近年，日常生活に必要となる情報を得る手段としてWebが注目されるようになり，Webが主要なメディアとして認知されるに至った．これは，実社会においてもWebの情報が影響力を持つということである．また，近年ユーザが容易に情報を発信することが可能なBlog（Weblog）やSNS（Social Networking Service）のようなユーザ参加型のコミュニティサイトが提供されることで，ユーザが任意のテーマに対して，日常生活の中で得た情報を発信するという点がメディアとしてのWebの特徴である．これは，任意のテーマに対して複数のユーザが評価情報を関連付ける仕組みのため，一方的な情報配信となる新聞，ラジオ，テレビのような従来型メディアよりも信頼性が高いと考えられる．

これらのコミュニティサイトは，ユーザ参加を前提としたメディアとなるため，CGM（Consumer Generated Media）と呼ばれる．しかし，ユーザが容易に情報を発信することが可能な結果，CGMでは様々な情報が氾濫し，評価情報もユーザによって異なる場合が多いために，CGMの閲覧者は，自身に必要な情報を取得する必要性から以前よりもメディアリテラシーが求められるという問題点が存在する．そして，CGMの中で近年着目されている情報が学校非公式サイトと呼ばれるWebサイトである．学校非公式サイトとは，学校が公式に運営するサイトとは別に開設され，主に中高生の間にて情報交換が行われる各種のコミュニティサイトを示す．

2008年1月～3月に行われた文部科学省による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の存在が確認され，これらのWebサイトでは，Blogや掲示板などのユーザの情報発信が容易な形式

*†文教大学

が多くを占める．さらに，学校非公式サイトの内容として誹謗・中傷などの表現が含まれるWebサイトはサンプリング調査により50%であることが確認されている．学校非公式サイトの利用率としては，閲覧したことがあるという回答が23.3%であり，これは学校運営においても影響力がある数値と判断できる．実際に学校非公式サイトから端を発したいじめなども発生していることから，教育現場においては，学校非公式サイトの動向に注意する必要がある．

このような経緯から，フィルタリングによって生徒や学生が学校非公式サイトを閲覧できないようにする取り組みも行われているが，フィルタリングの利用率は，神奈川県の教育委員会の調査によると，小学生で29.6%弱，中学生で22.6%，高校生で10.5%程度であり，必ずしも高い数値とは言えない．また，文部科学省の学校評価ガイドライン[2]を確認すると，学校教育に対して透明性を求めており，日常的な教育に関する評価情報が含まれることが多い学校非公式サイトから情報を遮断することは，透明性の確保に繋がらず，必ずしもよいとは考えられない．また，学校評価支援として，評価情報のデータ基盤を整備するというアプローチ[3]も行われているが，現時点でデータ基盤の整備を行っても，既にWeb上に氾濫している評価情報を活用することはできない．このため，学校教育の運営においては，常に学校非公式サイトの動向に注意し，学校非公式サイトに掲載される評価情報について，有用な情報については学校教育の改善に取り入れるといった必要性があると考えられるが，学校非公式サイトは，非公式という特性上，その存在が秘密裏であることが多く，頻繁なWebサイトの移動という問題がある．また，学校非公式サイトがCGMの場合には，情報の更新頻度が高いという点から，現在の主要な情報検索の手段である検索エンジンでリアルタイムに情報発信が行われている学校非公式サイトから情報を取得することは困難である．これは検索エンジンがCGMなどについては，類似情報の混在や情報量が膨大であるために，意図的にクローリングを回避しているという背景も影響している．このため，学校非公式サイトから継続的に情報取得を行うためには，学校非公式サイトの情報取得に特化した新たな情報検索を支援するサービスの早期提供が必要である．

本提案は，前述のような問題により，従来の検索システムを利用することでは，情報取得が困難である学校非公式サイトから日常的な情報の取得を支援することを目的とする．本提案を活用することで，教育現場に対して，継続的な学校評価情報の提供を行うサービスを創出するための手法を提示することを目的とする．これは現在着目されている学校非公式サイトからの情報抽出だけではなく，学術的な観点からも近年着目されているWebからの有用な動向情報を抽出するという試みであり，本提案は，他の類似問題に適用することもできると考えられる．

2. 研究の目的

本提案では，学校非公式サイトの情報検索を支援することを目的とするが，学校非公式サイトから情報を取得するためには以下の2つの問題点がある．最初の問題点は，学校非公式サイトでは前述の通り，頻繁なWebサイトの移動などにより，現在も活用されている学校非公式サイトを取得することは困難という点である．次の問題点は，Webの匿名性や運営を秘密裏に行うという学校非公式サイトの特性から，特定の学校の情報を取得することが困難という点である．

このため，前者の問題で，情報の検索時に課題となる具体的な問題点は，学校非公式サイトが頻繁に移動する点と，該当のWebサイトが学校非公式サイトであることを隠すためや具体的な学校名の判別を困難にするために，当て字表現や利用者の間で共通で利用される他の単語に言い換えを行った表現が用いられるという点である．本提案では，これらの問題に対応するために，情報の掲載時間に関する情報を取得することや情報の更新頻度を判定することで，取得した学校非公式サイトが現在も運用されていることを自動的に判定する．また，学校非公式サイトの特徴の1つである情報の発信媒体が通常の検索エンジンでは情報取得が困難な点に対応するために，通常のクローリング以外にBlogや掲示板に特化した情報取得方法を提供する．さらに，学校非公式サイトで用いられる固有の単語を認識するために，事前抽出した学校非公式サイトから，言語資源を蓄積することで，これらの表現を自動的に学習することで，Webのクローリングに学習した単語を活用することが可能となり，学校非公式サイトを効率的に発見する．

そして，後者の問題に対応するために，学校非公式サイトから固有表現となる情報の抽出を行う．ここでの固有表現とは，学校名だけではなく個人名や地方名などのWebサイトを実社会と関連付けるための情報を示す．しかし，学校非公式サイトのような秘密裏に運営されるサイトでは，先の問題と関連するが，個人特定の回避のために，学校非公式サイトにおいての固有の表現が用いられることが多い．これらの表現に対

しても，先の問題と同様に学習データから言語資源を蓄積することで対応を行う予定である．さらに，Webサイトに含まれる情報を解析することで，類似情報をグループ化し，利用者に対して，情報取得が容易なインターフェイスを提供する．これは，学校などの特定が困難な情報であっても，利用者に対して，関連の可能性がある学校非公式サイトの候補を提供することが可能になるために，従来の情報検索よりも有用であると考える．さらに，利用者の候補結果からの手動による判定を教師データとすることで，言語資源の更新を継続的に行うことも実現できると考える．

3. 処理の概要

3.1 処理の流れ

本提案による研究システムでは学校非公式サイトを自動的に抽出するための処理を概要図として図1に示す．

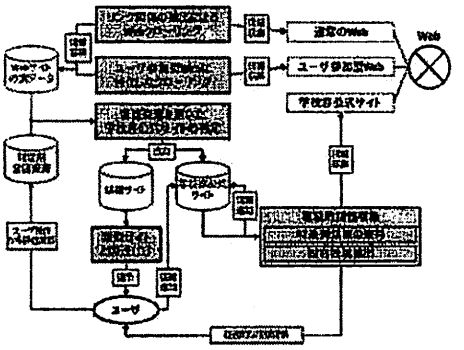

図1 処理の概要

Figure 1. Flow of Processing.

図1より処理部分のみを箇条書きで示すと以下の通りである．

- リンク関係の抽出によるWebクローリング
- ユーザ参加型Webに特化したクローリング
- 言語資源を用いた学校非公式サイトの判定
- 候補リストのグループ化と可視化
- 学校非公式サイトの時系列情報取得
- 学校非公式サイトの固有表現抽出

本提案による研究システムでは，上記の処理の中で，「リンク関係の抽出によるWebクローリング」，「ユーザ参加型Webに特化したクローリング」において，学校非公式サイトの判定候補となるWebサイトを取得する．そして，「言語資源を用いた学校非公式サイトの判定」において，取得したWebサイトが学校非公式サイトであるか否かを判定する．さらに，「候補リストのグループ化」，「学校非公式サイトの時系列情報取得」，「学校非公式サイトの固有表現抽出」において，学校非公式サイトから詳細な情報を抽出することで，学校教育の参考となる情報を取得することを目的としている．

3.2 情報のクローリング

「リンク関係の抽出によるWebクローリング」では，通常の検索エンジンと同様にリンク関係に着目したWebサイトの取得を行い，Webサイトの情報蓄積を行う．これは，Blogや掲示板以外の学校非公式サイトを取得するためである．具体的な方法としては，Webサイトからリンク情報をリストとして取得し，リンク対象となるWebサイトをクローリングするという再帰的な処理を行う．従来方式のWebクローリングでは，この再起処理において，複数のサイトからリンク関係にあるWebサイトを重要な基点として扱い，基点となるWebサイトからのリンクを重要度の高いリンクとする方式が主要である．しかし，現在の検索エンジンの一般的なクローリング手法だけでは，特定の種類の情報を保有するWebページ群を優先的に取得することは困難であるため，取得するWebページの種類に応じた専用のアルゴリズムを追加する方式が望ましい．本提案では，取得するWebページの対象が学校非公式サイトというリンクが好まれないWebサイトであるという特性を持つことから，クローリングを行う際には，リンク関係に対して，事前に学習を行った単語の重み付けを利用するなどで，クローリングを効率的に行う．また，Webサイトは，個人の嗜好に関する情報を考慮した形式[4]やサイト単位での判定を行うよりもコミュニティとして情報抽出を行う方式[5][6][7]やサイトの意味情報を利用する方式[8][9]が従来検索方式よりもよい精度を得られる．しかし，リンク関係の収集時にキーワードやコミュニティによる重み付けを行うような，情報収集を効率化する手法は，大規模なクローリング処理で適用するには，処理の負荷が大きいため，リンク箇所のテキストによりクローリングの重み付けを行うことで，コミュニティ解析などの代用とする．具体的な解析方法としては，後述の学校非公式サイトの特徴単語の固有値とリンク箇所のテキストを比較し，固有値の高い順から優先的にクローリングを行うアルゴリズムとする．

次に,「ユーザ参加型 Web に特化したクローリング」では,ユーザ発信型メディアなどの Web サイトの取得を行い,Web サイトの情報を蓄積する.これらのメディアでは,Blog の場合には情報検索を可能にするために,RSS (RDF Site Summary) などの情報が提供されるなどの特徴があるため,これを利用する.また,掲示板は,情報の更新頻度が Blog よりも頻繁であり,情報も検索に対応した形式になっていない.このため,Web サイトが掲示板などのユーザ参加型の Web サイトであると判定された場合は,優先的にクローリングを行う.ユーザ参加型の Web サイトの判定では,そこに掲載されるテキストに,時間情報を含む点や類似する HTML が連なる点からの判定とする.また,ユーザ参加型 Web サイトの自動判定は高度な処理となるために,判定が困難となる可能性がある.その場合には,主要なユーザ参加型 Web サイトに対して,検索機能を提供する URL と引数を本提案システムに手動で登録し,言語資源から学校非公式サイトで重要な単語とされる単語を取得し,これを検索キーワードとすることで,検索結果ページを基点とした情報のクローリングを行う.

3.3 非公式サイトの判定

「言語資源を用いた学校非公式サイトの判定」では,取得した Web サイトを複数のカテゴリに分類し,学校非公式サイトのカテゴリに含まれる Web サイトのグループを取得する.具体的な分類方法としては,式(1),式(2),式(3)に示す通り,Web サイトの重要度の高いキーワードを抽出し,それらの重要な単語を判定することで学校非公式サイトに含まれる可能性の高い単語と比較する.

$$TF(n) = \frac{\mathrm{Term(n)}}{\sum_n \mathrm{Term(n)}}. \quad (1)$$

$$IDF(n) = \log \frac{\sum_m \mathrm{Document(m)}}{df(\mathrm{Term(n)})}. \quad (2)$$

$$TFIDF(n) = TF(n) \times IDF(n). \quad (3)$$

式(1)では,単語の出現頻度となる TF 値を取得する.算出方法は,文書内の任意の単語を総単語で割ることにより,文書の長さや単語数に依存することのない出現頻度の算出を行う.また,式(2)では,単語の特徴となる IDF 値の算出を行う.算出方法は,任意の単語において,全ての文書数を任意の単語が出現する文書数で割った結果を対数化する.そして,式(3)では,単語の重要度を示す TFIDF 値の算出として,TF 値と IDF を掛けた値を算出する.ここで,IDF 値の算出で対数化を行う理由であるが,TFIDF 値の算出において,対数化を行わなかった場合には,IDF の影響が強くなり,出現頻度にあまり依存しない特徴抽出となるためである.そして,学校非公式サイトを形態素解析した結果として得られた全ての単語において,重要度抽出を行うことで,学校非公式サイトの特徴となる単語群を抽出する.重要な単語抽出後の補正としては,学校非公式サイトの特徴が,ユーザ参加型 Web サイトである可能性が高いために,実データには,絵文字や当て字表現などが非常に多く含まれる.このため,形態素解析の結果から記号やアルファベット,カタカナ表示,1 文字で構成される単語を削除し,Web サイトの特徴を強く示す名詞のみを抽出した.さらに,TF 値が一定以上である単語を排除し,IDF 値が任意の範囲内である単語群のみを抽出する.次に,学校非公式サイトの判定では,学校非公式サイトを学習データとして解析し,全ての単語において重要度計算を行った値を特徴ベクトルとして利用し,判定する Web サイトにも同様の処理を行う.これらの手順を式(4),式(5),式(6)に示す.

$$V_x = \{\mathrm{TFIDF}(x_1), \mathrm{TFIDF}(x_2), \ldots, \mathrm{TFIDF}(x_3)\}. \quad (4)$$

$$V_y = \{\mathrm{TFIDF}(y_1), \mathrm{TFIDF}(y_2), \ldots, \mathrm{TFIDF}(y_3)\}. \quad (5)$$

$$sim(V_x, V_y) = \frac{\sum_n (\mathrm{TFIDF}(x_n) + \mathrm{TFIDF}(y_n))}{\sqrt{\sum_n \mathrm{TFIDF}(x_n)^2} + \sqrt{\sum_n \mathrm{TFIDF}(y_n)^2}}. \quad (6)$$

式(4)は,学校非公式サイトの学習データから作成した各単語の特徴値をベクトルとしたものである.また,式(5)は,判定を行う Web サイトの各単語の特徴値をベクトルとしたものである.そして,式(6)において,x,y は,比較する Web サイトの集合とする.また,2 つのベクトル *Vx*, *Vy* を仮想空間上に配置し,座標を持たせ,両者のベクトルについてコサイン相関値である *sim(Vx,Vy)* を算出することで,類似判定を行う.

3.4 詳細情報の取得

「学校非公式サイトの時系列情報取得」では,Web サイトに含まれる時間を示す情報を取得する他に,各 Web サイトを追跡調査することにより情報の差分を取得することで,更新量や更新頻度を判定して現在も活用されているサイトであることを判定することを目的とする.特に時間に関する情報は,形態素解析の結果により,数値を表す名詞として,年月日を示す 8 桁の数値かもしくは年月日と時間を示す 12 桁の数値として抽出されるために,判定が容易である.そして,これらの数値を示す名詞は,更新頻度が高い場合には,

出現頻度を示すTF値が高くなり，任意の話題についての波及範囲が広い場合にはIDF値が低くなる傾向が見られる．Webサイトが現在も更新されているWebサイトであるかを判定する場合には，これらの数値も活用する．また，差分抽出には，Webサイトの特徴となる重要度をベクトルとして比較することで，大きくWebサイトの話題が変更された状況を認識することが可能である．

そして，「候補リストのグループ化と可視化」では，取得したWebサイトの中で，学校非公式サイトの可能性が高いが自動判定に至らなかった候補のリストをユーザに提供するが，提供する情報量が膨大となる可能性もあるために，通常の検索方式だけではなく，Webサイトの関係をグラフで表現したユーザインターフェイスを提供することで，ユーザの情報取得を支援することを目的とする．グラフ表現では，関連性の高いWebサイトを近くに配置し，距離によって意味の類似性を表現する．ユーザによるグラフ表現のインターフェイスを利用した履歴は，本研究システムに蓄積され，そこからユーザ特性や正誤情報を取得することにより，言語資源の更新を行うことで，動的な教師データの取得と言語資源の更新を行う．

また，「学校非公式サイトの固有表現抽出」では，Webサイトに含まれる固有表現を抽出し，情報のグループ化をより強固なものとする．さらに，それに類似するグループとして判定されたWebサイトは，類似Webサイトである可能性が高い集合体としてユーザに併せて提供する．固有表現の抽出は，日本語の場合には，形態素解析のような基本的な方法と組み合わせて文節情報を用いる方式[10]やSVM（Support Vector Machine）を用いる方式[11]が高精度の抽出を実現している．しかし，学校非公式サイトでは，固有表現の発見を困難にする要素として，当て字表現等があるために，これに対応するための言語資源を利用することで，精度の向上を行う．

4. 評価実験

4.1 検索キーワードの有効性検証

本提案システムでは，Webサイトのクローリングに際して，学校非公式サイトの特徴となる単語を用いることで，検索効率を向上させる．そして，この方式が有効と実証するために，匿名掲示板を対象として，学校非公式サイトにおいての重要語句を検索エンジン利用時の検索キーワードとして用いた場合と学校非公式サイト以外の学習データから抽出した重要語句を利用して，同様の方法にて検索を行った結果を比較し，実際の検索結果の上位50件において，学校非公式サイトおよび学校関連の話題を取り扱ったWebサイトが含まれる件数を比較した．ここで，検索結果に教育機関関連の話題を取り扱うWebサイトを含む理由は，本検証が学校非公式サイトのクローリング手法の検証であることから，教育機関関連の話題を取り扱ったWebサイトは，学校非公式サイトへのリンクが含まれる可能性やWebサイトの関係性をグラフで表現した場合に近いノードに存在する可能が高いと考えられるために，学校非公式サイトをクローリングする場合には重要な基点となるWebサイトと判断できるためである．本検証の結果を表1と表2に示す．

表1　重要語句を用いた検索結果

Table1. Search Result Using Keywords.

特徴単語	重要度	検索結果
実名	0.06204	12
攻撃	0.05777	4
周年	0.02099	4
竹早	0.02052	2
大泉	0.01854	0
モカ	0.01828	0
川崎	0.01774	2
引用	0.01553	1
経済	0.01545	1
学生	0.01519	15
二葉	0.01513	5
西南	0.01441	12
データ	0.01327	2
チッピー	0.01311	0
白百合	0.01288	24
稚い	0.01249	0
大学	0.01171	27
会長	0.01165	0
花火	0.01134	0
羞恥心	0.01051	0
関学	0.01037	15
同志社	0.01013	18
畳田	0.01009	0
葉平	0.01009	3
さくら	0.00994	0
合計件数（延べ件数）		147
平均件数		5.88

表2　重要語句を用いない検索結果

Table2. Search Result not Using Keywords.

単語	重要度	検索結果
ダリ	0.11149	0
利権	0.05337	0
偏見	0.05029	1
自民党	0.03200	0
追求	0.02831	0
エロス	0.02336	1
民主党	0.02127	0
投稿	0.02052	0
根源	0.01841	0
諸悪	0.01841	0
返信	0.01815	0
美術館	0.01572	0
タメ	0.01503	2
受験	0.01458	10
原因	0.01387	0
卒業生	0.01351	6
目先	0.01291	0
あたし	0.01223	2
卒業	0.01151	4
利益	0.01145	0
ニキビ	0.01128	0
医学	0.01092	1
出世	0.01008	1
てめぇ	0.01008	0
合計件数（延べ件数）		28
平均件数		1.12

表1と**表2**を確認すると特徴値と検索結果の件数については明確な相関を見出すことはできないが，**表1**においては，「西南」，「白百合」，「関学」，「同志社」などの固有表現や「大学」，「学生」などの明確に学校関連の話題に関連すると考えられる単語を除いた場合には，特徴値が高い単語ほど検索結果に学校非公式サイトや関連 Web サイトの件数が若干多いという弱い相関を見ることができる．また，**表2**の通常の言語資源から取得した単語を用いた場合の検索結果には，学校非公式サイトや関連 Web サイトが含まれる件数が一様に少ないことを確認できた．さらに，**表1**の結果の詳細を確認すると，全ての単語で平均して数値が高い訳ではなく，特定の単語において，学校非公式サイトや関連 Web サイト多く含まれるという傾向があり，さらにアルゴリズムを追加することで，これらの単語の絞込みを行う必要性を感じることができる．また，通常の Web サイトを対象としたクローリングでも同様の結果が得られると考えられるが，学校非公式サイトが匿名掲示板や Blog，SNS などの媒体を利用していることが多いために，本手法は，これらのユーザ参加型の Web サイトに適用することが効率的であると考えられる．

4.2　学校非公式サイトの特徴抽出

本提案の評価を行うために，手動にて取得した学校非公式サイト 43 件に対して，特徴となる単語の抽出を行った．具体的な方法として，学校非公式サイトから文字データのみを抽出し，形態素解析を行うことで，単語の出現頻度および特徴値を取得した．この際に，Web サイトを構成するスクリプトやスタイルなどの表現や操作性に関するデータは事前に削除を行った．そして，学校非公式サイトの Web ページ群において重要性の高いと算出された単語の上位 20 件を抽出した結果を**表3**に示す．

表3　学校非公式サイトの重要語句

Table3. Keywords of School Unofficial Website.

単語	出現頻度	特徴値	重要度
実名	0.06557	0.94612	0.06204
攻撃	0.06571	0.87918	0.05777
周年	0.01683	1.24715	0.02099
竹早	0.02169	0.94612	0.02052
大泉	0.01303	1.42325	0.01854
モカ	0.01466	1.24715	0.01828
川崎	0.01246	1.42325	0.01774
引用	0.01515	1.02531	0.01553
経済	0.01633	0.94612	0.01545
学生	0.02225	0.68288	0.01519
二葉	0.01063	1.42325	0.01513
西南	0.01284	1.12222	0.01441
データ	0.01183	1.12222	0.01327
チッピー	0.01051	1.24715	0.01311
白百合	0.00905	1.42325	0.01288
稜い	0.02046	0.61033	0.01249
大学	0.01617	0.72428	0.01171
会長	0.01136	1.02531	0.01165
花火	0.00797	1.42325	0.01134
羞恥心	0.00739	1.42325	0.01051

表 3 の結果を確認すると，「実名」や「攻撃」などのキーワードが特徴として挙がっていることを確認することができる．これらの単語は実際の教育現場において，いじめなどに関連する単語として出現することが多く，学校教育において，関連の深い重要語句が抽出されていると考えられる．また，表3の結果は，学校非公式サイトのグループにおいての特徴であるため，学校非公式サイトとそれ以外のサイトを判別するためには，学校非公式サイト以外のWebサイトを学習データとして追加し，その差分から両者の判定を行うことが有効であると考える．このため，本実験では，学校非公式サイト以外のデータを追加し，学校非公式サイトで重要語句として抽出された単語を除いたデータを表4に示す．

表4　学校非公式サイト以外の重要語句

Table4. Keywords other than School Unofficial Website.

単語	出現頻度	特徴値	重要度
ダリ	0.09935	1.12222	0.11149
利権	0.03750	1.42325	0.05337
偏見	0.03533	1.42325	0.05029
自民党	0.02248	1.42325	0.03200
追求	0.02270	1.24715	0.02831
エロス	0.01873	1.24715	0.02336
民主党	0.01706	1.24715	0.02127
投稿	0.03743	0.54818	0.02052
根源	0.01293	1.42325	0.01841
諸悪	0.01293	1.42325	0.01841
返信	0.03139	0.57815	0.01815
美術館	0.01105	1.42325	0.01572
タメ	0.02599	0.57815	0.01503
受験	0.02803	0.52016	0.01458
原因	0.01466	0.94612	0.01387
卒業生	0.02464	0.54818	0.01351
目先	0.00907	1.42325	0.01291
あたし	0.02351	0.52016	0.01223
卒業	0.01991	0.57815	0.01151
利益	0.00918	1.24715	0.01145

表4は，学習データとして，学校非公式サイト以外の教師データとして，Yahoo! Japanの各カテゴリから，学校非公式サイトと同様の体裁を持つユーザ参加型のWebサイトを取得し，重要と判定された単語の違いを確認するものである．

表3と表4のいずれの結果においても，重要度が0.05から0.1の範囲の少数の特徴が高いとされる単語と，0.01から0.03の2種類のグループに分割することができる．学校非公式サイトの結果に着目した場合には，特に重要とされる少数の単語は，先のWebクローリングでも学校非公式サイトの関連が高かったと考えられることから，これらの少数グループおよび固有名詞となる単語が学校非公式サイトの判定においても高い結果が得られるのではないかと予測される．

4.3 実験の考察

「検索キーワードの有効性検証」では，Webサイトをクローリングする際に，無作為なクローリングではなく，特定の単語に着目したクローリングを行うことで，学校非公式サイトの取得を効率化することが可能であるかを検証した．現在の学校非公式サイトは，との特性から，当て字や固有の表現を用いることで，学校非公式サイトと認識することが困難なように工夫がされている．しかし，本実験により，学校非公式サイトを実際に利用している利用者が用いる単語を活用したクローリングを行うことで，効率的にWebをクローリングするだけでなく，従来は発見が困難であった上記のような表現を用いたWebサイトを発見することができると考えられる．

また，「学校非公式サイトの特徴抽出」において，学校非公式サイトで用いられている特徴単語の抽出を行った．これは言語資源の蓄積だけでなく，実際の教育現場において，重要となる単語が含まれており，さらなる関連単語の抽出や学校非公式サイトの自動判定に有効に活用できるものと考えられる．しかし，今回の実験が小規模な解析であったことから，学習データを増加させることで，より明確な単語の特性を解析する必要があると考える．さらに，学校非公式サイトは，小学校，中学校，高等学校，大学により，話題が大きく異なると考えられるために，これらに共通して用いられる単語と各グループにおいてのみ特性となる単語を個別に検証する必要もあると考えられる．

5. おわりに

先の文部科学省の学校非公式サイトに関する調査で多数のサイトが認知されるに至ったが，これは氷山の一角に過ぎず，現在もその数は増加していると考えられる．また，現在の学校教育において，学校非公式サイトの対策としては，有害サイトとして扱い，情報のフィルタリングを行うという方式である．このような

方式は，悪意を持った学校非公式サイトには有効であると考えられるが，利用者が日常的な情報交換を行うWebサイトなどの本来は有効活用が可能である情報源を排除するものである．このため，これらのサイトから日常的に情報収集を行うことで，現在も情報交換が行われている学校非公式サイトから継続的な情報抽出を行うことは重要である．

そして，本提案を教育評価支援として適用することで，教育者は，生徒や学生の日常的な意見や疑問などをリアルタイムに認識することができるため，問題に対する即時の対応が可能となり，教育の品質を改善することができると考える．また，本提案は，学校教育の透明性確保にも貢献すると考えられるために，教育の信頼性確保にも繋がると考えられる．

今後の課題としては，学校非公式サイトの利用者が情報を交換する主要な手段として用いるのが携帯端末などであることから，情報収集においても携帯端末を用いることで，学校非公式サイトの利用者と同様の環境にて情報を取得することで，より現実の教育現場に活用が容易な方法を提供することが望ましいと考えられる．また，本提案は，現在の主流な方式として学校非公式サイトの対策に利用されているフィルタリング方式と組み合わせて適用することで，大きな効果を発揮することができると考えられる．このため，悪意のある学校非公式サイトと純粋な評価情報を掲載する学校非公式サイトを判定し，両方式を個別に適用することにより，さらなる学校教育の改善に繋げたいと考える．

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) 文部科学省：学校評価ガイドライン，
http://www.mext.go.jp/a_menu/shotou/gakko-hyoka/index.htm．

3) 久保裕也，玉村雅敏，木幡敬史，金子郁容：カスタマイズ可能な調査スキーマの共有による学校評価支援，情報処理学会論文誌，Vol.46，No.1，pp.172-186(2005)．

4) 堀田知宏，丸山崇，北栄輔：パーソナライズを考慮したWeb検索フィルタリングアルゴリズム，情報処理学会数理モデル化と問題解決研究会研究報告，Vol.2006，No.135，pp.89-92(2006)．

5) 丸山謙志，王冠超，徳山豪：Web検索結果におけるクラスタリングアルゴリズムの研究，情報処理学会アルゴリズム研究会研究報告，Vol.2005，No.26，pp.17-24(2005)．

6) 野村早恵子，小山聡，早水哲雄，石田亨：Webコミュニティ発見のためのHITSアルゴリズムの分析と改善，電子情報通信学会論文誌，Vol.J85-D-I，No.8，pp.741-750 (2002)．

7) 加藤一民，松尾啓志：Markov Cluster Algorithmを用いたWebコミュニティ群の発見手法，情報処理学会自然言語処理研究会研究報告，Vol.2005，No.22，pp.87-93(2005)．

8) 友部博教，松尾豊，武田英明，安田宗，橋田浩一，石塚満：Semantic Webのための人の社会ネットワーク抽出と利用，情報処理学会論文誌，Vol.46，No.6，pp.1470-1479(2005)．

9) 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，Vol.47，No.6，pp.1687-1695(2006)．

10) 山田寛康：Shift-Reduce法に基づく日本語固有表現抽出，情報処理学会自然言語処理研究会研究報告，Vol.2007，No.47，pp.13-18(2007)．

11) 中野桂吾，平井有三：日本語固有表現抽出における文節情報の利用，情報処理学会論文誌，Vol.45，No.3，pp.934-941(2004)．