ニコニコ動画における投稿者に特有な言葉の抽出
Extraction of Words Specific to Contributors in Niconico

上薗　竣
Shun Uezono
法政大学情報科学部ディジタルメディア学科
E-mail: shun.uezono.3j@stu.hosei.ac.jp

Abstract

Internet cultures have developed remarkably in recent years. For this reason, various specific words are born in individual community sites such as Niconico. Such specific words are difficult for beginners to understand because they are not familiar with the community sites. Therefore, it will be better for them to get able to easily recognize such specific words. However, many of such specific words are unknown words that do not appear in ordinary dictionaries. This paper proposes a method that extracts such specific, possibly unknown words from comments in Niconico. It extracts specific words by performing the following procedure. First, it extracts words of interest from comments given to contributors. Next, it calculates weight values of words associated with contributors, which determines specific words that do not appear widely. Finally, it shows relations among contributors and specific words by network visualization. This paper presents the results of experiments where subjects without such specific knowledge read visualization results, and shows that the method extracted specific words about contributors. However, it indicates a problem that the method sometimes erroneously extracts partial sentences or words as specific words.

1. はじめに

近年，インターネット文化の発達が著しい．それに伴って動画共有サイトであるニコニコ動画をはじめとしたコミュニティサイトで，独自のコミュニティが数多く形成されている．ニコニコ動画では，投稿者が独自の動画自由に投稿できる．また，個人で生放送配信が可能であるため，独自のコミュニティが形成されやすい．そのような独自のコミュニティの形成に伴って，その中でよく使われる特有な言葉が出現している．このような言葉や文化は，そのコミュニティを知らない人や馴染みのない初心者には理解することが難しい．また，そのような言葉の中には，一般的な言葉だけでなく，ネットスラングやアスキーアートといった特殊なものも数多く存在し，既存の辞書に含まれない場合も多々ある．

これまでにも，Twitter におけるツイートの分析のために，形態素解析を用いた未知語の抽出を行った研究 [1]が存在する．しかし，ニコニコ動画の場合は，付与されるコメントに一般的な言葉が用いられないことが多々あるため，形態素解析を行うことが難しい．本論文の著者が知る限り，ニコニコ動画を対象とした未知語の分析はまだ行われていない．

本研究では，ニコニコ動画投稿者コミュニティにおける動画へのコメントから，形態素解析を用いずに未知語を含む特有な言葉の抽出を行う手法を提案する．文字列の切り出しを行い，文字列同士を比較し評価する．そして，似たような言葉を省略したのち，最終的に抽出された特有な言葉を可視化する．まず，コメントから N-gram 法を用いて文字列を切り出し，他投稿者の動画に付与されるコメントの文字列と比較を行う．投稿者ごとに特有な言葉を抽出するために，TF-IDF 法を用いて特定の投稿者のみに頻出する語と，数多くの投稿者の動画に汎用的に用いられる言葉を判定する．この際，似たような言葉は，レーベンシュタイン距離を用いて類似度を計り省略する．最後に，抽出された特有な言葉をネットワーク可視化によって提示する．

提案手法を評価するために，独自のコミュニティを形成する投稿者を選定し，抽出された言葉と投稿者の関係をネットワーク可視化した例を示す．更に，コミュニティの知識がない被験者によるアンケートに基づく実験の結果を示す．結果として，その投稿者のみに見られる顔文字や，辞書に見られないような単語や文章などを抽出することができた．ただし，抽出される特有な言葉の一部には，文章や単語の一部を切り出した言葉が抽出されることがあった．すなわち，本来一つの文章や単語であるが，文字列を切り出したものが独立して特有な言葉として抽出されてしまうことがあった．そのような場合，その言葉だけでは意味を理解することが難しいという問題点がある．

2. 関連研究

形態素解析が難しい言葉がコメントとして広く用いられるニコニコ動画では，未知語の抽出を行う研究は，本論文の著者の知る限りまだ行われていない．山田の研究 [1]では，形態素解析に基づき Twitterにおける未知語を抽出する．解析精度を高めるために，単語辞書の拡張及び，文字種と出現パターンのルールを設定した未知語抽出モデルの適用し，形態素解析の最適化をしている．ツイート分析に伴う辞書の新語，流行語の補充には，はてなキ

Supervisor: Prof. Hiroshi Hosobe

ーワードを用いている．辞書化が難しい言葉が数多く用いられるニコニコ動画では言葉の辞書登録が難しいため、形態素解析を適用することができない．そこで，本研究では辞書を用いない手法により未知語を含んだ特有な言葉の抽出手法を提案する．

森らの研究 [2]では，N-gram 法を用いて，自然言語の解析段階で必要とする未知語に対処するための単語を抽出する手法と，それらの品詞を推定し辞書の語彙を増強させる手法を提案している．宮崎ら [3]は，話題判定の評価値の決定に TF-IDF 法を用いている．TF-IDF 法に基づいて有用なキーワードを選択する手法の提案をしている．また，赤塚ら [4]は，類似度を計る手法にレーベンシュタイン距離を用いている．類似度を計り，多数アラームシステムを類似度により適正化する手法を提案している．

3. 準備

3.1. ニコニコデータセット

ニコニコデータセット[1]は国立情報学研究所により公開されているニコニコ動画の動画情報及びコメントデータである．ニコニコ動画に 2012 年 11 月初旬までに投稿された約 830 万件の動画のメタデータと，それに対するコメントデータが付属される．本研究では，動画に対するコメントデータを用いる．

3.2. ニコニコ動画 API

ニコニコ動画 API はニコニコ動画の公式 API である．動画 ID，動画タイトル，投稿者のユーザ ID などの動画情報を直接サーバーから取得することができる．本研究では，ニコニコデータセットに付随しない情報である投稿者の情報を利用するために API を用いる．

3.3. N-gram 法

N-gram 法は文字列を検索する手法である．文章に対して一定の文字数で文字列を切り出し，検索を行う．これによって，辞書や構文解析を必要としない文字列探索を可能とする．本研究では，コメントから頻出する文字列を抽出するために，N-gram 法を用いる．

3.4. レーベンシュタイン距離

レーベンシュタイン距離 [5]は，文字列の類似度を計る手法である．二つの文字列を同じ文字列にするために，何回の挿入，削除の編集が必要かを計ることによって，文字列同士の類似度を計る．本研究では，切り出された文字列群に対して，似たような文字列を省略する．そこで，レーベンシュタイン距離を用いて，距離の近い文字列を検索するために用いる．

3.5. TF-IDF 法

TF-IDF 法は特徴量を計ることで文字列の重みを決定する手法である．ある文書における文字列の出現頻度である TF 値と，文字列が出現する文書頻度の逆数の対数である IDF 値を掛け合わせた，TF-IDF という特徴量を用いて，特徴語を抽出する手法である．文字列iの文書jにおける出現回数を$n_{i,j}$，文書jに出現する文字列の総数を$\sum_k n_{k,j}$とし，総文書数をD，文字列iが出現する文書の数をdf_iとすると，文字列iの文書jにおける TF 値$\mathrm{tf}_{i,j}$，IDF 値idf_i，TF-IDF 値$\mathrm{tfidf}_{i,j}$は以下の数式で表すことができる．

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \qquad \mathrm{idf}_i = \log\frac{D}{\mathrm{df}_i}, \qquad \mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i$$

本研究では，投稿者に付与されるコメントから切り出した文字列に対して，特有であるかを計るための重みづけをするために TF-IDF 法を用いる．

4. 提案手法

本研究では，投稿者に特有な言葉を抽出するための手法を提案する．最初に，投稿者に特有な言葉を抽出する．その後，投稿者に特有な言葉を可視化する．

4.1. 投稿者に特有な言葉の抽出

投稿者の動画を調査し，それらに付与されたコメントを利用して，未知語を含む投稿者に特有な言葉を抽出する．対象とする投稿者を選定し，動画のコメントから N-gram 法を用いて文字列を切り出す．その後，各投稿者におけるそれらの文字列の重みを TF-IDF 法により評価する．また，抽出される特有な言葉の中で似たような言葉を，レーベンシュタイン距離を用いて識別し，抽出される言葉の一部を省略する．個々の処理の詳細を以下に述べる．

4.1.1. 投稿者コミュニティの検索

投稿者のコミュニティを検索により特定する．投稿者の動画をニコニコ動画 API とニコニコデータセットの動画情報データを照合することで検索する．その後，それらの動画に対するコメント情報をニコニコデータセットのコメントデータから入手する．

4.1.2. N-gram 法による文字列切り出し

動画に付与されたコメントから N-gram 法を用いて文字列の切り出しを行う．コメント文を 3 文字，5 文字というように，数文字単位に分割し判定に用いる文字列を切り出す．そして，投稿者ごとに頻出する文字列を記録する．図 1 は，実際にコメントから切り出される文字列と，その頻度を示した例である．

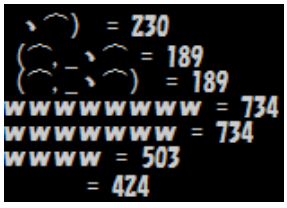

図 1．切り出される文字列とその頻度の例

4.1.3. TF-IDF 法による特有な言葉の抽出

TF-IDF 法に基づいて，投稿者ごとに頻出する文字列と他投稿者に頻出する文字を比較する．ある投稿者から抽出された文字列に対して，その投稿者に対して現れた回数を TF 値とする．また，複数の投稿者に対して幅広く使用されている文字列は投稿者にとって特有ではないとし，

[1] http://www.nii.ac.jp/dsc/idr/nico/nico.html

ある文字列の投稿者全体での出現回数から IDF 値を計算する．二つの値から，投稿者ごとに抽出された文字列の TFIDF 値を計算し，閾値を超えたものを投稿者にとって特有な言葉として抽出する．

4.1.4. 類似する言葉の省略

レーベンシュタイン距離を用いて，類似する言葉を省略する．N-gram 法によって切り出される文字列は，辞書による抽出を行わないため，同じような文字列の言葉が数多く抽出される．そこで，TF-IDF 法を行い抽出された投稿者全体の特有な言葉の中から，似たような文字列の言葉を省略する．レーベンシュタイン距離を用いて，特有な言葉同士の距離を計り，距離が近い言葉を省略する．この際，それぞれの投稿者で，TF-IDF の値が小さく，重みの低いものを省略する．図 2 は距離が近い言葉を検索し，省略をした結果の例である．

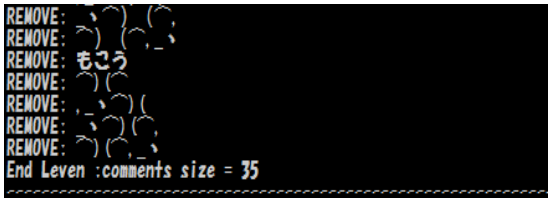

図 2. レーベンシュタイン距離を用いた言葉の省略

4.2. 投稿者に特有な言葉の可視化

抽出した言葉を最終的にネットワーク可視化によって提示する．可視化のツールとしてCytoscapeを用いる．可視化では，投稿者と，投稿者それぞれから抽出された特有な言葉の関連を表したネットワークの集合を形成する．可視化によってコミュニティにどのような特有な言葉が用いられているかを明確に判別できるようにする．また，共通の言葉が抽出された投稿者同士を隣接させて表示することで，抽出される言葉と投稿者だけでなく，近いコミュニティを有する投稿者同士の関係も視覚的に認識可能とする．

投稿者と特有な言葉との関係を可視化した例を図 3 に示す．投稿者を水色の枠で表示し，それぞれの投稿者から湧出された特有な言葉を青色の枠で表示している．

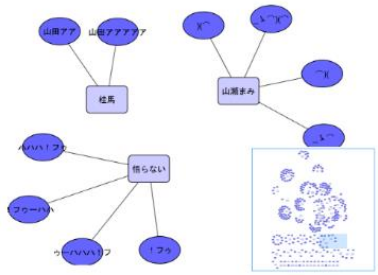

図 3. 特有な言葉の可視化

5. 実装

ニコニコデータセット及び，ニコニコ動画 API を用いて，提案手法に基づく特有な言葉を抽出するシステムを Processing環境にて実装した．

5.1. 投稿者の検索

ニコニコデータセット及びニコニコ動画 API を利用して，実験対象とする投稿者を選別する．API から投稿者の情報を獲得し，再生数やコメントの情報をデータセットの情報と照らし合わせ対象とする投稿者を決定する．

5.2. 投稿者に特有な言葉の抽出

ある投稿者の動画に付与されるコメントに N-gram 法を適用し文字列の切り出しを行う．本研究では切り出す文字列の長さを 3 文字から 8 文字とした．そして，投稿者に対して 100 回以上頻出した文字列を頻出する文字列として記録する．TF-IDF 法を用いて対象とする投稿者全てに対して現れる文字列と他投稿者に現れる文字列と比較する閾値を 0.03 とし，これを満たす文字列を特有の言葉として抽出する．次に，似たような言葉を省略するために，投稿者全体の特有な言葉からレーベンシュタイン距離を適用する．距離が 2 以下であると判定された言葉は，TF-IDF 値がより小さいものを省略する．最終的に投稿者ごとに特有な言葉を抽出する．

6. 実験

提案手法に基づき，特有な言葉の抽出と可視化を行う．再生回数 100000 以上の動画を投稿したことがある投稿者 2500 名を，独自コミュニティを持つ可能性のある投稿者として実験の対象とする．

6.1. 抽出結果

条件を元に抽出される全体のネットワークは以下の通りである．図 4 のネットワークは，各投稿者で抽出された特有な言葉から，結果をより明確にするために複数の投稿者に共通して現れる特有な言葉を省略したものである．省略される対象の例を図 5 に示す．同じ言葉が使われていた投稿者同士は隣接して表示されている．また，共通な言葉が一切用いられていない投稿者は独立したネットワークを形成している．

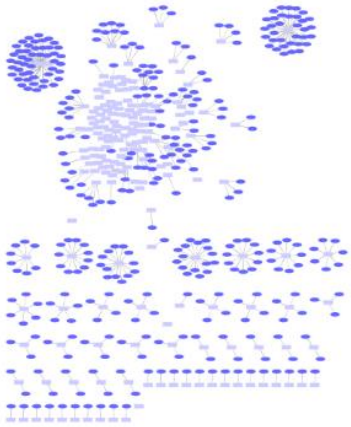

図 4. 特有な言葉のネットワーク可視化

評価実験では共通に表れる特有な言葉を省略した．図 5 における黄色の範囲が複数の投稿者に共通して現れる言葉であり，図 4 で省略されて表示されている．

図 5. 複数の投稿者に現れる特有な言葉の例

ネットワーク可視化された投稿者と特有の言葉の集合からは，その投稿者にのみ付与された言葉を認識することができた．また，顔文字といった，通常の辞書では識別することのできない言葉も抽出することができた．また，特有な言葉として抽出される言葉には投稿者の名前を含むものや，固有名詞，感嘆詞を含む言葉が数多く抽出された (図 6) .

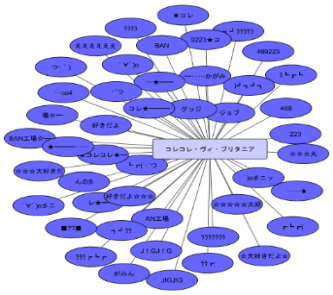

図 6. 投稿者の特有な言葉のネットワーク可視化の例

一方で，記号のみの文字列を切り出したものなど，言葉としての意味がない文字列や，言葉の意味が理解できない文字列も一部特有な言葉として抽出されてしまうことがあった． また，ある一つの文章として意味を持つ言葉が分割されて抽出されるという問題点もあった(図 7) .

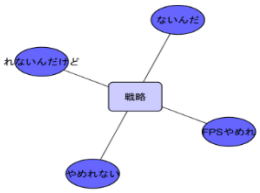

図 7. 本来一つの文章が分割されて抽出される例

6.2. 評価実験

知識のない複数人に対して可視化したネットワークから得られる特有な言葉が妥当であるかの評価実験を行う．評価方法としては，ニコニコ動画及び，投稿者についての知識がない 10 名の被験者に対して図 4 を提示する．ネットワーク図から特有な言葉が読み取ることができるかを「1. 読み取ることができる」，「2. 概ね読み取ることができる」，「3. あまり読み取ることができない」，「4. 殆ど読み取ることができない」の 4 段階の評価を行ってもらい，提案手法の妥当性を検証する．

評価実験の結果を表 1 に示す．結果として，ある程度の投稿者に対する特有な言葉をネットワークから読み取ることができるという評価を得た．一方で，問題点として，文章の一部から切り出された言葉から，一部の投稿者について特有な言葉の理解が難しいという意見が挙げられた．

表 1. 被験者による評価実験の結果

評価	人数
1. 読み取ることができる.	1
2. 概ね読み取ることができる.	5
3. あまり読み取ることができない.	4
4. 殆ど読み取ることができない.	0

7. 議論

本研究では，形態素解析を用いない言葉の抽出を行った．結果として，投稿者にとって特有であるという特徴をもった言葉の抽出を可能とした．また，顔文字や，感嘆詞含む言葉など幅広い言葉が抽出された．一方で， 辞書による形態素解析を行わないことの問題点も存在する．頻出する文字列を抽出するにあたり，第一に，本来一つの文章や言葉が分割されて別の特有な言葉として抽出されることがある．第二に，顔文字として使われている記号の一部が抽出されるなど，必ず意味の読み取れる言葉が抽出されるとは限らないという問題点がある．TF-IDF による特徴量のみではこのような言葉も特有な言葉として抽出してしまう．対策として，独自の辞書を汎用的な文字列から独自の辞書を作成することが考えられる．辞書からの形態素解析を取り入れることで，更に特有な言葉の認知精度の向上に繋がると考えられる．

8. おわりに

本研究では，ニコニコ動画のコミュニティに対して言葉の抽出を試みた．そして，辞書を用いることができない場合でも未知語の抽出を行う手法の提案を行った．可視化の結果及び評価実験から，認識可能な妥当性を持った特有な言葉の抽出及び判別を行うことができた．しかし，意味が判別できない言葉が抽出される課題も存在する．本研究の拡張として，対象とするコミュニティで，汎用的に出現する言葉の抽出から更に独自の辞書を作成する．そして，形態素解析も行うことで，より高い精度で抽出を行うことができると考える．

文　献

[1] 山田勉, "Twitter 分析のための形態素解析の最適化," 言語処理学会第 20 回年次大会, pp. 578-581, 2014.

[2] 森信介，長尾眞, "n グラム統計によるコーパスからの未知語抽出," 情報処理学会論文誌, vol. 39, no. 7, pp. 2093-2100, 1998.

[3] 宮崎将隆, 川端豪, "Wkipedia ページへの tfidf 法の適用," 情報処理学会研究報告：音声言語情報処理 (SLP), vol. 77, no. 2, pp. 1-6, 2009.

[4] 赤塚祥太，野田賢，杉本謙二, "レーベンシュタイン距離に基づく連鎖アラームの類似性解析," 化学工学論文集, vol. 39, no. 4, pp. 352-358, 2013.

[5] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710, 1966.