動画共有サイトにおける動画ナビゲーションのためのコメント要約手法

松原　宏和† 　新妻　弘崇†† 　太田　学†

†, †† 岡山大学大学院自然科学研究科　〒 700–8530 岡山県岡山市北区津島中 3-1-1
E-mail: †{matsubara, ohta}@de.cs.okayama-u.ac.jp，††niitsuma@suri.cs.okayama-u.ac.jp

あらまし　YouTube やニコニコ動画といった有名な動画共有サイトには日々膨大な動画が投稿されており，動画内容の簡潔な要約は視聴者にとって大変有益である．そこで本稿では，動画閲覧支援のためのコメント要約手法を提案する．提案するコメント要約の特徴は，コメントから抽出した単語にアプリオリアルゴリズムを適用して重要単語を抽出し，特徴的な感想表現により分かり易く感想を可視化する点にある．実験により感想表現の抽出性能と重要コメント抽出性能などについて評価する．

キーワード　動画ナビゲーション，ニコニコ動画，コメント，要約

A comment summarization method for navigating to a video-sharing site

Hirokazu MATSUBARA†, Hirotaka NIITSUMA††, and Manabu OHTA†

†, †† Graduate School of Natural Science and Technology, Okayama University
3-1-1, Tsushima-naka, Kita-ku, Okayama, 700–8530 Japan
E-mail: †{matsubara, ohta}@de.cs.okayama-u.ac.jp，††niitsuma@suri.cs.okayama-u.ac.jp

Key words　Video navigation，Niconico，Comment，Summarization

1. はじめに

近年，YouTube [1] やニコニコ動画 [2] などの動画投稿サイトによる動画共有が盛んになっている．このような動画共有サイトは誰もが自由に動画を投稿できるため，膨大な数の動画が存在する．例えばニコニコ動画では，2013 年 6 月時点で一般会員登録者が約 3,215 万人，有料会員は約 200 万人で，全動画投稿数は，2013 年 2 月 4 日時点で 2,000 万件を突破した [3]．ニコニコ動画のコメントの特徴としては，動画再生中にコメントを書き込むことが出来る点が挙げられる．このコメントは，コメント投稿時刻の先頭からの動画再生時刻を保持しており，その動画再生時刻が来るとコメントが表示される．そのためコメントには，動画の各再生時刻での視聴者の感想や動画内容について書かれている．また，コメントに書き込める文字数は最大 75 文字であり，視聴者は動画を見ている途中で書き込むため，視聴者の感想や動画の内容について，短い言葉で書かれているものが多いことも特徴の一つである．ユーザがコメントを投稿するのは，特定のタイミングで大量のコメントを投稿する「弾幕作成」,「あいさつ」,「疑問があったとき」,「相槌」,「驚いたとき」,「感動したとき」,「ツッコミをいれたいとき」,「一緒に歌いたいとき」など閲覧者の感情が動かされるときに多いという研究結果がある [4]．また，動画の内容を端的にあらわしているものとして動画の説明文があるが，動画の説明文やタイトルは動画の投稿者が書いたものであり，必ずしも動画の内容と合致しないことがある．そのため動画によっては視聴するまで，どのような動画であるのかを把握することが難しいことがある．そこで重要なコメントを抽出し，要約してユーザに提示すれば内容把握の支援となると考えた．

本研究では，コメントの中から重要なものを選択し，選択したコメントから感想に特徴的な表現を抽出して可視化する手法を提案する．重要なコメントの抽出は，まずコメント中の単語に着目して，アプリオリアルゴリズム [5] を適用して単語の相関ルールを求める．次に求めたルールの中で支持度，確信度が上位のものを重要単語とし，その単語を含むコメントを抽出する．そして，感想に特徴的な表現により，抽出したコメントを動画の感想を述べたコメントと，それ以外のコメントに分けて可視化する．本研究でいう感想に特徴的な表現とは，視聴者の感想が含まれている評価表現と叫喚表現のことである．評価表現は「おもしろい」,「好きだ」のように形容詞や形容動詞のことである．叫喚表現は,「〜きたあああああ」や「うわあああああああああああああ」など母音を繰り返すような強い感情を表現したものを利用して抽出した,「きたあ」や「うわあ」などの表現である．本研究では，この「〜きたあああああ」や「うわあああああああああああああ」のような母音を繰り返すコメントを叫喚コメントと呼ぶ．

本稿の構成は次の通りである．2. 節で関連研究について述べ，

3. 節で相関ルールに基づく重要コメント抽出について説明し，4. 節で感想表現抽出によるコメントの分類とコメント要約ついて説明する．5. 節で重要コメント抽出と，コメントの分類実験について述べ，6. 節でまとめる．

2. 関 連 研 究

2.1 動画ナビゲーション

動画のサムネイルに吹き出しを付与し，そこに要約を表示してニコニコ動画のナビゲーションを行うシステムが提案されている [6]．このシステムでは，一度視聴したことのある映像コンテンツの中で印象に残ったシーンを再び見返す場合に，効率よくそのシーンを見つけることができる．また，コメント数の多い箇所とユーザの印象に残ったシーンの箇所は，一致することが多いという傾向が報告されている．この調査結果に基いて，コメント数の多いシーンからサムネイル画像を取り出す．具体的には，まず取り出した画像をコメント数が多いほど大きいサイズとなるように表示して並べる．次に，各サムネイルに最頻出コメントを吹き出しとして付与し，一覧できるようにする．コメントの全文をそのまま吹き出しとして付与すると，ぱっとみて分かるサムネイルとしての機能を果たせなくなる恐れがあるため，コメントの最初の 5 文字のみを吹き出しとして付与している．しかし，コメント数が多いシーンがユーザの印象に残るシーンであるとは限らないため，コメント数が多いが印象に残らないシーンを取り除き，コメント数がそれほど多くないが印象に残るシーンを追加することを，今後の課題としてあげている．

2.2 重要文抽出

複数の Web ページから重要文とキーワードを抽出することにより要約を提示する方法が提案されている [7]．キーワードは，文書全体の形態素解析の結果，名詞となった単語の中から TF-IDF 法を用いて重要度を計算して抽出している．重要文抽出では，まずキーワードの重要度を利用して，文書重要度を計算する．次に，文書重要度が高い文書からキーワードの重要度を利用して文重要度を計算し，文重要度が高い文を抽出する．また，一定長以下の文は含まれる情報が極端に少なく，重要文としては不適切であるため，20 文字以下の文は重要度が低くなるようにしている．

2.3 意見文抽出

大学などの教育機関におけるレポートでは，課題に対するレポート提出者の意見が含まれていることが多い．そのためレポートからの意見文の抽出手法の提案がされている. [8] の研究ではレポートの各文の文末表現に注目し，表 1 に示す語で文末が終わっており，かつ 3 つ以上の名詞を含む文を意見文として抽出した．表 1 の単語は，一般的な文末表現の中から，レポートにおいて意見が述べられている可能性が高いと考えられる表現が選ばれている．本研究では，感想を表す文を抽出するために，文中の単語の品詞や叫喚表現を利用するが，このような文末表現は考慮しない．

SVM と新聞記事を用いた Weblog からの意見文抽出手法も提案されている [9]．この研究では，SVM を用いて記事を主観的な意見を含むレビュー記事と非レビュー記事に分類し，新聞記事から抽出した特徴語を用いて作成した辞書に基づいてレビュー記事から意見文を抽出する．レビュー記事からの意見文抽出には，訓練フェーズとテストフェーズがあり，それぞれのデータを形態素解析器 MeCab にかけ形態素解析する．訓練フェーズでは，新聞記事の社説に含まれる文を全て意見文，国際のジャンルに含まれる文を全て非意見文と仮定し，これらの文から評価表現となる可能性が高い形容詞，形容動詞，動詞の形態素を取り出し特徴語リストを作成する．作成した特徴語リストを基に，以下の式で特徴語にスコアリングし，特徴語辞書を作成する．

表 1　意見文の文末表現 [8]

Opinion Expression	Modalities
考え	思う，考える，感じる，言える
推量	う，かも
要望	たい，望ましい
自発	れる，られる
主張	べき，ちがいない

$$score(w_i) = \frac{P_O(w_i) - P_F(w_i)}{P_O(w_i) + P_F(w_i) + k} \quad (-1 \leqq score(w_i) \leqq 1)$$

ここで $P_O(w_i)$ は意見文に特徴語 w_i が出現する確率で，$P_F(w_i)$ は非意見文に特徴語 w_i が出現する確率である．k は経験的に 0.001 としている．特徴語辞書を用いて，形態素解析を行ったテストデータの文にスコアを付ける．得られた文のスコアに基づき，文を意見文と非意見文に分類する．文の意見文と非意見文への分類には以下の式を用いる．

$$\text{Score}(s) = \sum_{w_i \in s} score(w_i)$$

文 s に含まれる特徴語のスコアの総和 Score(s) が 0 より大きければ意見文，0 以下であれば非意見文と分類する．ただし文が短い場合など，特徴語が 1 語も現れない場合は，スコアが 0 となり非意見文に分類される．

3. 相関ルールに基づく重要コメント抽出

ここでは，アプリオリアルゴリズムを利用した重要コメント抽出について説明する．まず 3.1.1 節で説明する方法でコメントの各文から，特徴語を抽出し，特徴語リストを作成する．この特徴語リストに対して，アプリオリアルゴリズムを適用して，得られた相関ルール，支持度，確信度を基に重要コメントを決定する方法について 3.2 節で説明する．

3.1 特徴語リストの作成

3.1.1 コメントからの特徴語抽出

ここでは，重要コメント抽出のためにまずコメントから特徴語を抽出する方法について説明する．

最初にコメントのローマ字を全て半角に変換し，MeCab を利用して形態素に分割する．MeCab の解析結果は，表層形，品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用形，活用型，原形，読み，発音に分かれている．本研究では，このうち

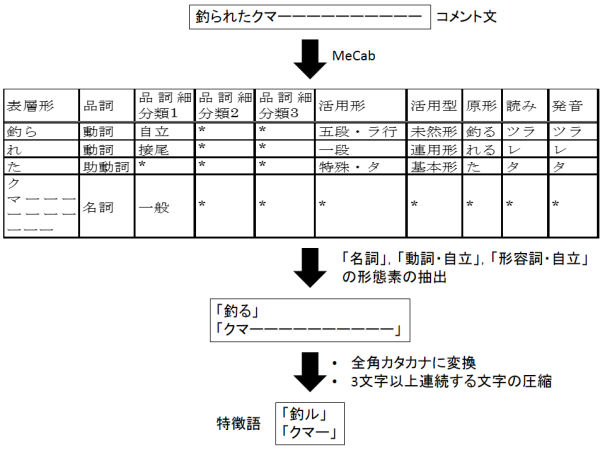

図 1　特徴語抽出処理の例

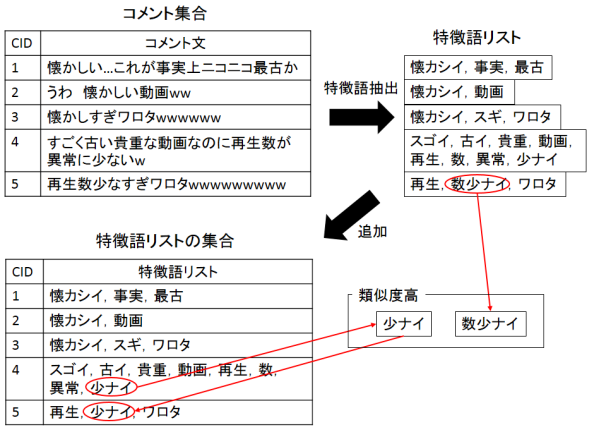

図 2　特徴語集合の作成例

品詞，品詞細分類 1，原形を用いる．分割した形態素が「名詞」であれば，ひらがな，半角カタカナを全て全角カタカナに変換し，さらに 3 文字以上連続する同じ文字を 1 文字に圧縮する．形態素が「動詞」または「形容詞」であれば，品詞細分類 1 が「自立」の形態素のみを残し，その品詞の原形を全角カタカナに変換し，さらに 3 文字以上連続する同じ文字を 1 文字に圧縮する．図 1 にこのような特徴語抽出の例を示す．

ここで形態素を全角カタカナに変換する理由として，同じものを指しているが微妙に違った書き方で書かれるものを同一特徴語として扱うためである．例えば「くま」や「クマ」,「クマ」を半角カタカナで書いている場合は同じものとして扱う．3 文字以上連続する文字を 1 文字に圧縮する理由は,「クマーーーー」や「クマーーーーーーーー」など,「ー」の数が違うだけで同じもの指している単語を同一特徴語として扱うためである．これにより「クマーーーー」,「クマーーーーーーーー」を「クマー」にできる．

3.1.2　特徴語リストと特徴語リストの集合の作成

次に 3.1.1 節で抽出した特徴語から，特徴語リストと特徴語リストの集合を作成する方法を説明する．まず，図 2 のようなコメント集合があれば，CID（Comment ID）が 1 のコメントから特徴語を抽出する．抽出した特徴語は，複数ある場合があるため，それらから特徴語リストを作成する．次に作成した特徴語リストを特徴語リストの集合に追加する．CID が 2 以降のコメントも同様に特徴語を抽出し，特徴語リストを作成する．この特徴語リスト中の特徴語と，特徴語リストの集合中の特徴語の類似度を Jaro-Winkler 距離を用いて計算する．この類似度が閾値より高ければ，特徴語リスト中の特徴語を特徴語リストの集合中の特徴語に置き換えた特徴語リストを，特徴語リストの集合に追加する．類似度が閾値より低ければ，特徴語リストを特徴語リストの集合に加える．

3.1.3　Jaro-Winkler 距離

Jaro-Winkler 距離は二つの文字列の類似性を測る指標である．これは 0 以上 1 以下の値をとり，1 が完全一致を表す．Jaro-Winkler 距離は以下のようにして計算する．まず 2 つの文字列 s_1, s_2 が与えられたとき，Jaro 距離 d_j を求める [11]．d_j は

$$d_j = \begin{cases} 0 & \text{if m = 0} \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

となる．m は s_1 と s_2 で一致した文字の数である．ただし文字が

$$\left\lfloor \frac{max(|s_1|, |s_2|)}{2} \right\rfloor - 1$$

以上離れている場合は一致した文字とみなさない．t は異なる順序で一致した文字の数を 2 で割ったものである．例えば文字列「MARTHA」と「MARHTA」であれば，全ての文字が一致するが，T と H の一致する順序が異なるため $t = 2/2 = 1$ となる．文字列「DWAYNE」と「DUANE」であれば，D，A，N，E の文字が一致し，一致する順序が同じため，$t = 0$ となる．次に，Jaro-Winkler 距離 d_w は

$$d_w = d_j + \left(l * p\left(1 - d_j\right)\right)$$

で定義される [12]．l は 2 つの文字列の先頭から一致する文字の数である．p は定数で，通常 $p = 0.1$ が使われる．

2 つの文字列の類似度を測る手法として他にも Levenshtein 距離があるが，Levenshtein 距離は 2 つの文字列の長さが短い場合，2 つの文字列の類似度が大きくならない．例えば「クマー」と「クマ」に Jaro-Winkler 距離を用いれば，類似度は 0.911 であるが，Levenshtein 距離から類似度を計算すれば，0.667 となる．本研究では形態素のように短い文字列でも類似度を大きくしたいため，Jaro-Winkler 距離を用いる．

3.2　相関ルールに基づく重要コメント抽出

3.1 節で説明した特徴語リストの集合にアプリオリアルゴリズムを適用し，得られる相関ルールから重要コメントを抽出する方法を説明する．3.2.1 節でアプリオリアルゴリズムについて説明し，3.2.2 節で重要コメントの抽出方法について説明する．

3.2.1　アプリオリアルゴリズムによる特徴語の相関ルール

アプリオリアルゴリズムは，Agrawal と Srikant [5] により提案され，トランザクションセットから，ある基準を満たした相関ルールを全て抽出する．全アイテム集合を $I = \{i_1, i_2, \cdots, i_m\}$ とし，その部分集合をアイテムセットと呼ぶ．D を全トラン

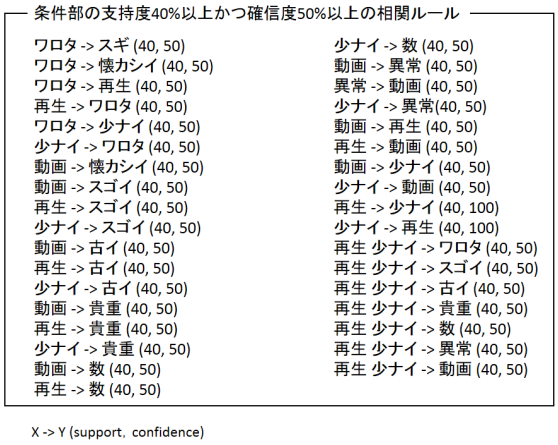

図 3　相関ルール抽出例

ザクションの集合とする．各トランザクション T はアイテムセットである．また $X, Y \neq \emptyset$ で $X, Y \subset I$ かつ $X \cap Y = \emptyset$ が成り立つとき，$X \rightarrow Y$ を相関ルールと呼び，X を条件部，Y を帰結部と呼ぶ．ルールの重要度を表わす指標として，支持度（support）と確信度（confidence）がある．アイテムセット X の支持度 $sup(X)$ は，D 中の X を含むトランザクションの割合である．ルール $X \rightarrow Y$ の支持度 $sup(X \rightarrow Y)$ は，$sup(X \cup Y)$ のことで，D 中の X と Y を共に含むトランザクションの割合である．また確信度 $conf(X \rightarrow Y)$ は $sup(X \cup Y)/sup(X)$ で定義される．

本研究では各トランザクション T を特徴語リストとし，全トランザクションの集合 D を特徴語リストの集合とする．図 2 のような特徴語リストの集合にアプリオリアルゴリズムを適用し得られる相関ルールの例を，図 3 に示す．図 3 では，条件部の支持度が 40% 以上かつ確信度が 50% 以上の相関ルールのみを示している．

相関ルールマイニングでは，同時に出現するアイテム同士の関係を相関ルールとして抽出することができる．特に，同時に出現する頻度が高いアイテムを見つけることが出来る．本研究では同じような内容のコメントが多く投稿されていれば，それはその動画の特徴をあらわすコメントであると考える．そこで，コメントから特徴語を抽出し，同時に使用される頻度が高い特徴語を見つけることで，重要コメントを抽出する．アプリオリアルゴリズムは，ある支持度，確信度以上の相関ルールを発見するアルゴリズムである．そのため本研究では，使用頻度の高い特徴語を見つけるために，アプリオリアルゴリズムを用いた．

3.2.2　重要コメント抽出

得られた全ての相関ルールを利用して重要コメントを抽出する．すなわち，得られた個々の一つの相関ルールの条件部と帰結部の特徴語を両方含むコメントを，重要コメントとして全て抽出する．

4.　要約コメントの可視化

3.2.2 節の方法で抽出した重要コメントを要約し，可視化する．まず重要コメントから，重要コメントを代表するコメントを抽出し，それを要約する．次に，全コメントから感想に特徴的な表現を抽出し，それに基づいて代表コメントを感想とその他に分類する．最後に，分類したコメントを可視化する．

表 2　ラベルコメントのソート結果の例

相関ルール	ラベルコメント
クマ -> ツエ (10.5, 15.5)	クマつええ w
クマ-> 釣ル (10.5, 12.1)	釣られクマー
クマ-> スゲル (10.5, 3.5)	クマのすげぇらしい
最古-> 聞ク (7.9, 25.0)	最古と聞き
最古-> ユーザー (7.9, 18.1)	ユーザー最古動画
最古-> ニコ (7.9, 4.6)	ニコ動最古の動画
スゲル-> コメント (3.6, 20.0)	コメントすげー
スゲル->ェナ (3.6, 10.0)	すげぇなコメント
スゲル-> ウゴク (3.6, 5.0)	うごきすげ www

4.1 節で重要コメントからの代表コメントとラベルコメントの抽出方法について述べ，4.2.1 節で感想に特徴的な表現の抽出方法について説明し，4.2.2 節で代表コメントの分類方法について説明する．4.3 節でコメントの可視化方法を説明する．

4.1　代表コメントとラベルコメントの抽出

3.2.2 節で抽出した重要コメント群からそれらの代表のコメントを抽出する．すなわち，それぞれの相関ルールを含む重要コメント群の中で異なる形態素数が最多のものを，その相関ルールを表す代表コメントとする．例えば,「くますっげえええええええええええ」と「クマの動きすげえええええ ww」という重要コメント群がある.「くますっげえええええええええええ」の異なる形態素数は，名詞の「くま」，動詞の「すっ」，名詞の「げ」，フィラーの「え」，感動詞の「ええ」の 5 になる.「クマの動きすげえええええ ww」の異なる形態素数は，名詞の「クマ」，助詞の「の」，名詞の「動き」，動詞の「すげる」，フィラーの「え」，名詞の「ww」の 6 になる．そのためこの例では,「クマの動きすげえええええ ww」が代表コメントとなる．また文字列長が最短のものを，その相関ルールのラベルコメントとする．ただし，相関ルールの条件部と帰結部の特徴語から抽出したコメントが一つのみであった場合は，そのコメントは代表コメントかつ，ラベルコメントとなる．次に，代表コメントとラベルコメントをそれぞれ，代表コメントとラベルコメントの抽出に利用した相関ルールの条件部の支持度により降順にソートする．さらに，同じ支持度の条件部を持つ相関ルールを，確信度により降順でソートすることで，その相関ルールの代表コメント，ラベルコメントをソートする．ソート後の相関ルールとそのラベルコメントの例を表 2 に示す．ここでは，条件部が「クマ」である相関ルールの条件部の支持度が 10.5 で最も高い．さらに，条件部が「クマ」である相関ルールの中では,「クマ-> ツエ」という相関ルールが最も確信度が高い．

4.2　感想に特徴的な表現によるコメントの分類

4.1 節で示した代表コメントを可視化する前に，感想に特徴的な表現によりコメントを感想とその他に分類する．4.2.1 節で感想に特徴的な表現の抽出方法について述べ，4.2.2 節でコメントの分類方法について説明する．

4.2.1 感想に特徴的な表現の抽出

感想に特徴的な表現は，評価表現や「すごいいいいいい」のような語尾の母音を繰り返す叫喚コメントから抽出する叫喚表現のこととする．なお,「すごいいいいいい」という叫喚コメントがあれば，これから抽出される叫喚表現は「すごい」となる．まず，本研究で用いる評価表現について述べ，続いて叫喚表現の抽出方法について説明する．

本研究では評価表現は，品詞が「形容詞」,「名詞-形容動詞語幹」のものとする．

全コメントからの叫喚表現の抽出は [13] の研究を参考にした．図 4 に叫喚表現の抽出例を示す．叫喚表現は以下のように抽出する．

- 動画の全コメント中の日本語の母音の小文字を大文字に変換する．
- コメントから正規表現を用いて叫喚コメントを抽出する．
- 叫喚コメントから繰り返し母音部分と繰り返し母音部分以前の文字列を抽出する．
- 繰り返し母音部分以前の文字列の最後の形態素を抽出する．
- 最後の形態素が「っ」や「ょ」など小文字で始まっている場合，最後の形態素の一つ前の形態素を最後の形態素の前に結合する．
- 最後の形態素が「っ」や「ょ」で始まっていない場合．
  - 最後の形態素が「名詞-一般」,「形容詞」,「動詞-自立」のいずれかであれば，その形態素と繰り返し母音部分を一文字にしたものを結合する．
  - 最後の形態素が「名詞-一般」,「形容詞」,「動詞-自立」のいずれでもなければ，それ以前の形態素を見て,「名詞-一般」,「形容詞」,「動詞-自立」のいずれかがあれば，その形態素から最後の形態素までを結合し，さらに繰り返し母音部分を一文字にしたものを結合する．
- 繰り返し母音部分以前の文字列に「名詞-一般」,「形容詞」,「動詞-自立」のいずれもない場合は，繰り返し母音部分以前の文字列と繰り返し母音部分を一文字に変換したものを結合する．
- 結合された文字列を叫喚表現とする．

叫喚コメントの抽出に用いる正規表現は

([あ – おア – オ])\\1{2,}

である．

図 4 であれば，繰り返し母音部分は「えええええ」で，繰り返し母音部分を一文字に変換したものは「え」である．そして，繰り返し母音部分以前の「これはｗｗｗｗすげ」の最後の形態素は「すげ」である．形態素解析結果「すげ」は,「動詞-自立」となる．したがって,「すげ」と「え」を結合し，叫喚表現は「すげえ」になる．ただし「あああああああああ」のような同じ母音のみで構成されているものは叫喚コメントとして抽出しない．

4.2.2 代表コメントとラベルコメントの分類

4.1 節で説明した代表コメントには，動画の感想が含まれて

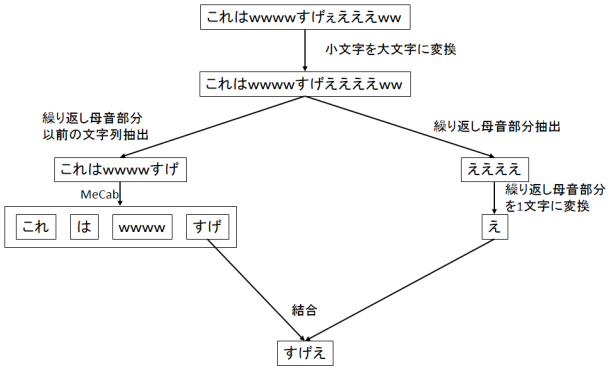

図 4　叫喚表現の抽出例

いると考えられる．そこで，代表コメントとラベルコメントを代表コメント中の感想表現の有無によって，感想コメントとそれ以外に分類する．代表コメントに感想表現が含まれていれば，それは感想コメントとする．

具体的にはまず，代表コメントを評価表現により分類する．代表コメントを MeCab を用いて形態素に分割する．その形態素の品詞が,「形容詞」,「名詞-形容動詞語幹」であればその代表コメントとラベルコメントを感想コメントに分類する．次に形態素に「形容詞」,「名詞-形容動詞語幹」のものがなければ，叫喚表現により代表コメントを分類する．全コメントから抽出した叫喚表現と代表コメントが部分一致すれば，その代表コメントとラベルコメントを感想コメントに分類する．感想コメントに分類されなかったコメントは非感想コメントとする．

4.3 コメントの可視化

4.1 節の方法でソートした代表コメントを順に可視化する．可視化例を図 5 に示す．ここでは，相関ルールの条件部の支持度の上位 5 件，さらに，同じ支持度の相関ルールの確信度の上位 5 件までを表示した．まず，代表コメントを 4.2.2 節の方法で感想コメントと非感想コメントに分類する．次に，分類された代表コメントをそれぞれ 4.1 節の方法でソートする．さらに，感想コメント，非感想コメントのそれぞれで，同じ条件部の支持度でかつ条件部が同じ相関ルールを一つのクラスと考える．ただし，条件部の支持度が同じで，条件部が異なるものは，別のクラスとする．そのクラス名は相関ルールの条件部とする．図 5 の，感想コメントでは，相関ルールの条件部の支持度が最も高いものが，条件部が「クマ」の相関ルールで，その次が「動画」を条件部として持つ相関ルールである．また各クラスの右側には，そのクラスの最も確信度の高い代表コメントを表示する．左側には，そのクラスのラベルコメントを確信度が高いものを上から順に表示する．

5. 実　　験

4.1 節で説明した方法により抽出した代表コメント，ラベルコメントが動画のメタデータとして有益な情報であるかどうかを評価するための実験を行う．実験では，国立情報学研究所の

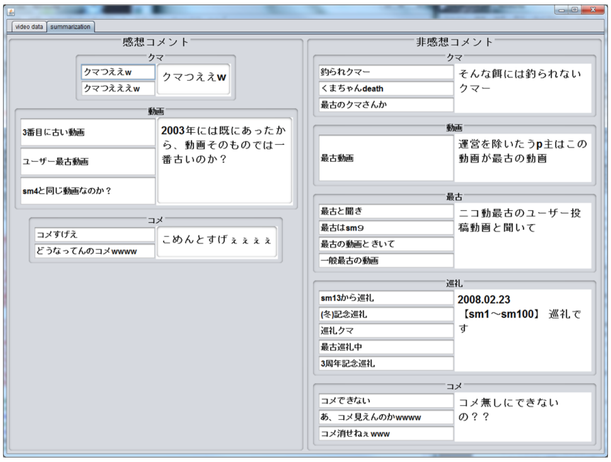

図 5　コメントの可視化例

表 3　実験に利用する動画データ

動画 ID	タイトル	カテゴリ	コメント数
sm14	クマとの死闘	動物	5,661
sm3190026	魔女の宅急便に登場の，かぼちゃとニシンのパイを作ってみた	料理	40,877
sm6264033	世界のすごい人詰め合わせ。	スポーツ	22,503
sm8155813	【作業用ＢＧＭ】 切ないピアノ曲を集めてみた	音楽	24,232
sm15822708	【実況】ポケットモンスターコイキング～金鯱の逆鱗～第 1 話	ゲーム	204,102

表 4　代表コメントの抽出例

相関ルール	重要コメント数	代表コメント
ソウ-> ウマイ (16.4, 51.0)	1,079	映画を見た当時はあんまりうまそうな 気はしなかったが、なかなかどうしてうまそうではないか (´ｐ`)
イイ-> イウ (6.9, 2.8)	25	かわいいいいいいいいい！ 100 点！この時点で 100 点 !!
イワシ-> イル (6.6, 17.1)	145	イワシいらなかったと思うがカボチャ&ニシンのパイだもんなぁ見ためバリいい w
イワシ-> ソウ (6.6, 4.8)	41	イワシは臭みが強そう代用する魚なら「スズキ」「鯛」がいいんでは？

表 5　代表コメントの比較

4.1 節の方法	最長のものを抽出する方法
かわいいいいいいいいい！ 100 点！この時点で 100 点 !!	いいい
普通にうまそうだけどこれ嫌いなのよね	クッソ普通にウマそう wwwwwwwww
シェ～フ　左上の杏仁豆腐は何ですか？	杏仁豆腐 wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww

ダウンロードサービスにより株式会社ドワンゴの「ニコニコ動画コメント等データ」を利用した [14]．表 3 に実験に利用した動画データについてまとめる．まず，5.1 節で代表コメント，ラベルコメントの抽出実験について述べ，5.2 節で感想表現によるコメント分類の実験について述べる．

5.1　代表コメントの抽出実験

4.1 節の方法で抽出した代表コメントやラベルコメントの妥当性を評価する．まず，相関ルールから抽出した重要コメントとその代表コメントについて分析する．本研究では，コメント中の異なる形態素数が最多のものを代表コメントとして可視化することで，動画内容把握の支援になると考える．動画 ID「sm3190026」のコメントから抽出した代表コメントの一部を表 4 に示す．この動画は，料理カテゴリの動画であり，映画「魔女の宅急便」に登場する「かぼちゃとニシンのパイ」を実際に調理する様子を撮影したものである．

表 4 より，代表コメントは，動画中に登場する料理を見た視聴者の感想や，調理者へのアドバイスなどがあることがわかる．ここで，代表コメントの妥当性を評価するために，重要コメントから代表コメントを以下のように抽出する．代表コメント群の中で文字列長が最長のものを代表コメントとして抽出する．この方法で表 4 と同じ動画から代表コメントを抽出した．この方法と 4.1 節の方法で抽出した代表コメントの例を表 5 に示す．表 5 の一行目の例は，相関ルール「イイ-> イウ」の代表コメントである．文字列長が最長のものを代表コメントとする方法では，「いいい」が代表コメントとなる．しかし，これは「い」の繰り返しで文字列長が長くなっているだけで，コメントの意味が良く分からない．4.1 節の方法であれば,「かわいいいいいいいいい！ 100 点！この時点で 100 点 !! 」が代表コメントとなる．この場合,「い」を繰り返しているコメントより，4.1 節の方法の代表コメントの方が，可視化するコメントとしては適切である．また表 5 の三行目の文字列長が最長のものを代表コメントとする方法では，ニコニコ動画でよく見られる「w」を多く使っているため文字列長が長くなっている．しかし，4.1 節の方法では,「シェ～フ　左上の杏仁豆腐は何ですか？」となる．「杏仁豆腐ｗｗｗ」だけでは,「杏仁豆腐」がおそらく動画中に出現するのだろうとわかるが,「左上の杏仁豆腐は何ですか？」であれば動画の左上に「杏仁豆腐」が出現することがわかる．実際に動画を見ると，この動画のメインの料理が出来上がったシーンの左上に杏仁豆腐が映っていた．よって，4.1 節の方法は,「wwww」や「いいいい」など同じ文字を繰り返すコメントが重要コメントに含まれる場合に有効である．

5.2　感想表現によるコメント分類実験

4.2.1 節で定義した感想表現を用いてコメントを感想コメントと非感想コメントに分類する．本研究で利用する感想表現には評価表現と叫喚表現がある．そこで，評価表現を用いてコメ

ントを分類した場合と，叫喚表現を用いてコメントを分類した場合のコメントの違いを実験により確かめる．

まず各動画の全コメントから叫喚表現を抽出する．次に，全コメントから評価表現と叫喚表現を含むコメントを感想コメントに，それ以外を非感想コメントに分類する．その分類結果を表 6 に示す．表 6 の感想コメントの計は，評価表現を含むコメントと叫喚表現を含むコメントの重複を除いた数である．非感想コメントは感想コメントに分類されなかったコメント数であり，合計は各動画の全コメント数である．この結果から動画によって感想コメントと非感想コメントの割合は異なることがわかる．表 6 の動画であれば，コメント中の感想コメントの割合は，最大で約 29%，最小で約 8%である．

動画 ID「sm3190026」のコメントの分類結果を見ると，「うまそううう」，「いいね」，「食べたい」，「焼き上がりが楽しみだあ〜」などの動画中で調理している料理に関する感想を表すコメントが確認できる．また，「あぶなっかしいいいいい」，「魚の捌き方おかしいぞー w 見てて怖い www」といった動画のワンシーンに関する感想を表すコメントも確認できた．

動画 ID「sm3190026」の動画に注目し，感想表現を用いてコメントを感想コメントとその他に正しく分類できるかどうかを評価した．対象は，表 3 の動画 ID「sm3190026」のコメント 40,877 件である．このコメントから，著者が人手で感想を述べているコメントと，その他のコメントを調べた．コメントを，感想表現を用いて感想コメントと非感想コメントに分類する．比較対象として，評価表現のみを用いてコメントを分類する方法と，叫喚表現のみを用いてコメントを分類する方法を評価した．表 7，8，9 にそれぞれ感想表現，評価表現，叫喚表現を用いてコメントを分類した結果を示す．なお，表中の「全体」は，全てのコメントの分類結果を示す．

表 7，8，9 より，感想表現を用いた方法が最も全体の精度が高いことがわかる．そして感想コメントへの分類結果では，感想表現，評価表現，叫喚表現のいずれも高い値を示しているため，感想表現を用いることで感想コメントを得られることがわかる．しかし，非感想コメントへの分類結果は，高いとはいえない．これは，感想コメントに分類すべきコメントを，非感想コメントに分類してしまっているからである．非感想コメントに誤分類したコメントを見ると，「キタ━━━━━ヽ(∀)━━━━━!!!!」や「sugeeeeeeeeeee」などがある．これらの例は，同じ記号やローマ字の繰り返しを叫喚コメントとして抽出すれば分類できると考える．しかし，コメント中に，「形容詞」や「形容動詞」を含まないが感想や意見を述べているコメントを分類できない．そのため，今回の方法で感想コメントに分類できなかったコメントを分類するための方法が必要である．

5.3 コメントの可視化実験

表 3 の動画のコメントを可視化した結果を図 6，7，8，9 に示す．ただし，動画 ID「sm14」の可視化結果は，図 5 と同じため省略する．図 5 以外の全ての動画で，非感想コメントより感想コメントの方が多く可視化された．

図 6 では，感想コメントに「うまそう」というコメントが目立つ．しかし，条件部の特徴語が異なるためこれらは別のクラスとして扱われる．これらをまとめて要約して可視化することが今後の課題として挙げられる．

表 6　動画の全コメント分類結果

動画 ID	感想コメント			非感想コメント	合計
	評価表現	叫喚表現	計		
sm14	321	219	480	5,181	5,661
sm3190026	11,906	7,743	1,4979	25,898	40,877
sm6264033	2,596	4,375	6,399	16,104	22,503
sm8155813	4,372	3,001	6,446	17,786	24,232
sm15822708	18,600	26,256	38,757	165,345	204,102

表 7　感想表現を用いたコメントの分類結果

	精度	(正解/分類数)
感想	0.956	(14,320/14,979)
非感想	0.670	(17,345/25,898)
全体	0.775	(31,665/40,877)

表 8　評価表現を用いたコメントの分類結果

	精度	(正解/分類数)
感想	0.966	(11,500/11,906)
非感想	0.607	(17,598/28,971)
全体	0.712	(29,098/40,877)

表 9　叫喚表現を用いたコメントの分類結果

	精度	(正解/分類数)
感想	0.954	(7,389/7,743)
非感想	0.533	(17,650/33,134)
全体	0.613	(25,039/40,877)

図 7 はタイトルが「世界のすごい人　詰め合わせ。」であるため，可視化したコメントにも「すげえ」といったコメントが上位に入っている．また「すげえ」などはユーザの感想であるため，感想コメントに分類されていることがわかる．

図 8 の動画は，様々なジャンルの切ないピアノの曲を流すものである．可視化したコメントを見ると，「いい曲」や「切なくなる」といった感想がある．

図 9 の動画はゲームの実況動画で，現在全 10 話あるが，この動画は第 1 話である．まず感想コメントの「イイ」クラスの代表コメントに「いい最終回だった」とあるが，この動画は最終回ではない．このコメントはニコニコ動画でよく使用されるもので，高いクオリティで最終回のように盛り上がり，非常に良い出来栄えの動画を賞賛するときに使用される．

6. ま と め

本研究では，コメント中の単語にアプリオリアルゴリズムを適用し，得られた相関ルールから重要コメントを抽出する手法を提案した．さらに，感想表現を用いてコメントを感想とその他に分類して，分かり易く提示する方法を検討した．

代表コメント抽出実験では，重要コメントから代表コメントを抽出し，重要コメント中で最も文字列長が長いものより，異なる形態素の数が最大のものを代表コメントとする方法が良いということがわかった．また，感想表現によるコメントの分類

図 6　sm3190026 のコメントの可視化結果

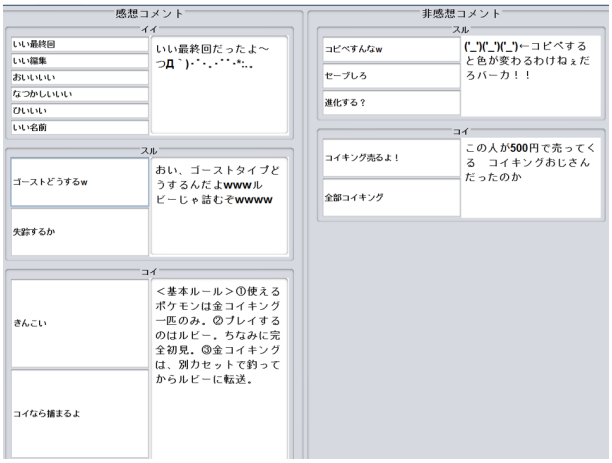

図 9　sm15822708 のコメントの可視化結果

図 7　sm6264033 のコメントの可視化結果

図 8　sm8155813 のコメントの可視化結果

実験も行った．その結果，評価表現や叫喚表現を用いることで，感想コメントを得られることが分かった．コメントの可視化では，五つの動画のコメントを可視化した．

今後の課題として，相関ルールの条件部の支持度以外での要約方法の検討や感想に特徴的な表現の追加が挙げられる．そのために，得られた相関ルールやその代表コメント，ラベルコメントについてさらに詳しく分析する．

文　　献

[1] YouTube
http://www.youtube.com/

[2] ニコニコ動画
http://www.nicovideo.jp/

[3] Wikipedia
http://ja.wikipedia.org/wiki/ニコニコ動画

[4] 青木秀憲，宮下芳明，“ニコニコ動画における映像要約とサビ検出の試み”，情報処理学会研究報告 2008-HCI-128/2008-MUS-75, Vol.2008，No.50，pp. 37-42，2008.

[5] Agarwal，Rakesh and Srilant，Ramakrishnan and others “Fast algorithms for mining association rules”，Proc. 20th Int. Conf. Very Latge Data Bases，pp. 487-499，1994.

[6] 中村貴洋，青木秀憲，宮下芳明，“マンガ的手法を用いたニコニコ動画ナビゲーション”，情報処理学会研究報告 2008-HCI-130, Vol.10(4)，pp. 103-110，2008.

[7] 柴田裕子，山内和子，石川知里，高田雅美，城和貴，“複数 Web ページの重要文抽出および直感的理解を支援するための GUI の開発”，情報処理学会研究報告．BIO，バイオ情報学，vol. 2007(128), pp. 81-84，2007.

[8] 砂山渡，川口俊明，田村幸寛，“レポートの課題との関連度と意見文抽出による情報量評価”，電子情報通信学会論文誌.D，情報・システム J93-D(10)，pp. 2032-2041，2010.

[9] 川口敏広，松井藤五郎，大和田勇人，“SVM と新聞記事を用いた Weblog からの意見文抽出”，人工知能学会第 20 回全国大会（JSAI2006），2006.

[10] MeCab：Yet Another PartOfSpeech and Morphological Analyzer, http://mecab.sourceforge.net/.

[11] Jaro，M.，“Probabilistic Linkage of Large Public Health Data File”，Statistics in Medicine，Vol. 14，pp. 491-498，1995.

[12] Winkler，W.，“String Comparator Metrics and Enhanced Decision Rules in the Fellegi-sunter Model of Record Linkage”，In proceedings of the section on Survey Research Methods，1990.

[13] 浅井洋樹，秋岡明香，山名早人，“きたあああああああああああああ!!!!! 11: マイクロブログを用いた教師なし叫喚フレーズ抽出”，DEIM Forum 2013 A4-4，2013.

[14] 大学共同利用機関法人情報・システム研究機構国立情報学研究所 (NII)：情報学研究データリポジトリ
http://www.nii.ac.jp/cscenter/idr/nico/nico.html