4V-01

Youtube のコメント欄を利用した動画概要の抽出

関塚　拡†　　秋岡　明香‡
明治大学 総合数理学部

1. はじめに

ネット上で情報を得たり,趣味娯楽を充実させたりするために使う媒体として YouTube が利用される場面が多くなってきている. 日々多くの動画が更新されていく中で同ジャンルの動画でも投稿者ごとに動画系統が異なることは多く,探している情報や好みの投稿者を見つけることは困難である.特にミュージックビデオや歌ってみた動画は,企画動画やゲーム実況動画などによく見られるタイトルやサムネイル画像に動画情報を集約するケースが少なく,好みの歌い手の動画にたどり着きづらい.

そこで動画を視聴せず得られる情報の一つであるコメント欄に注目し,それらをトピック分類や共起語の関係性などで解析し得た情報を組み合わせることにより動画概要を抽出できるのではないかと考えた.本稿では音楽系の動画に対して,各コメントに付随しているいいね数とコメントの内容を関連付けた SLDA[1]によるトピック分類を使用して動画の概要を整理することを検討する.これにより, 歌い手が違う同じ曲の動画に対しても容易に比較ができるようになり,情報収集の幅を広げることを目的とする.

2. 提案手法

本研究では YouTube API を用いて動画タイトルとコメント欄の取得を行いそれらの情報を使って SLDA で得られた結果と組み合わせて情報を整理していく. コメントに関しては一番先頭のメインコメントのみを取得しそのコメントに対しての返信コメントは重要でないと判断し取得対象外とした.主に日本語のコメントを拾い形態素解析エンジンの Mecab を使って「名詞,形容詞,副詞」に該当する単語を抽出し次の単語ごとの自作辞書定義や SLDA などの入力用にリスト型に格納し準備しておく. SLDA は自然言語処理で広く使われている潜在的ディリクレ配分（LDA）のトピック分類手法[2]をベースにさらに文章に紐づく数値の影響度も考慮したモデルである.LDA との違いは文書に付随した教師データに基づくトピックで分類する点である. 通常の LDA では一つのコメントに対してどういったトピックが潜在的に存在しているかを単語の出現確率で予測していたが,SLDA では加えて各コメントのいいね数という教師データを目的関数,そのコメントを生成しているトピックの割合を説明変数としてトピックごとのいいね数への影響度の予測を行っている.以下この影響度の数値をいいね係数と呼ぶ.これにより今回の YouTube のコメントに対してではどのトピックがいいね数に大きく関わっているかがわかりトピック自体に対する優先度がつけやすくなるのが大きな利点である.

まず自作辞書を用いてコメント全体における出現頻度が高すぎる単語や低すぎる単語を調整していく.出現頻度が１回のみの単語に対しては本来 LDA などのトピック分類をする場合には除外することが多いのだが,ほかのトピック分類が活用されるニュース記事やレビューと違い,一文が比較的短くかつ量も一つの動画内となると数が稼げないことが多いことから一定条件のもと採用した.その条件は出現回数が１回の単語が含まれているコメントに付随するいいね数が上位２５％以上のいいね数を稼げているコメントには出現頻度が１回でも辞書登録に含めるというものである.以上の条件のもと残ったコメントを１コメントずついいね数とセットでコーパスに登録していき, 一つの動画につき 500 回学習させていく.

今回出現頻度が上位１５位までの単語が共起の一つになっている場合は優先的に出力した.また出現頻度が多くない単語群でも３つ以上まとまった共起が見られる単語群なら数値の優劣のないサブトピックとして出力するようにした.また SLDA で得られた分類結果に関してはいいね係数が一番高いトピックのみを取り出した.トピックを構成している単語はそのトピック内での予測出現確率と対になっている.これらを用いて共起語の単語群に対するいいね指数をトピックの優先順位を作るために以下のように定めた.

$$likepoint(t) = x * \sum_{w \in t} p(w)$$

$likepoint(t)$: *単語群* t *についてのいいね指数*
x : *取り出したトピックのいいね係数*
$\sum_{w \in t} p(w)$: *単語群* t *の中で*
取り出したトピック内にある単語の出現確率の和

以上の流れで得られた結果を使って動画概要を表す.

Extract video summary using YouTube comment.
† Sekizuka Hiro, Meiji University
‡ Sayaka Akioka, Meiji University

3. 実験と結果

今回動画概要抽出に選んだ動画は THE FIRST TAKE による鈴木雅之の「違う,そうじゃない」[3]である.曲名である「違う,そうじゃない」は Mecab で形態素解析すると分割してしまうので「曲名」として指定してある.また鈴木雅之のほかの呼び名(鈴木さん,雅之さん,マーチンなど)は自作辞書に組み込む際にすべて鈴木雅之としてカウントするように指定した.動画概要として出力した結果を以下に示す.

表 1. 視聴者がいいねしているトピック

単語群	いいね指数
凄い―レベル―歌唱力	5.022
歌―うまい―上手―演奏	3.851
かっこいい―渋い―素敵	3.725
綺麗―関係	3.599
鈴木雅之―パフォーマンス―最近	3.369
最高―楽しい―ファーストテイク	3.034
歌声―色気―見た目	2.783
声―昔―全然―歳	2.783
良い―格好	2.218

表 2. 視聴者がコメントしているサブトピック

単語群
ドラム―かわいい―猫ちゃん―写真―猫
take―first―First
ラブソング―王様―帝王
大型―新人―アニソン
かっこ―世過ぎ―ええ
ダイアン―津田―笑
ネタ―画像―ネット
初めて―フル―曲名
そうじゃない―何
マーチン―ライブ
音源―ロ―CD
金―木―雅之

表 3. コメント欄全体での単語の出現頻度

単語	出現頻度
歌	107
かっこいい	99
良い	94
綺麗	90
声	85
鈴木雅之	79
最高	75
凄い	70
歌声	70
マーチン	55
初めて	54
ネタ	53
そうじゃない	50
好き	46
上手い	45

表 1 にある SLDA と共起語の整理情報に関しては,SLDA で得られた結果と組み合わせることのできた単語群を見るとこの動画の視聴者が多くいいねしているカテゴリが具体的に取り出せている.表 3 の単純な出現頻度の表と比較すると,「歌声-色気」や「鈴木雅之-パフォーマンス」など出現頻度だけでは取り出せない情報を取り出せていることがわかる.またサブトピックのほうも動画内のどういったシーンに視聴者が注目しているのかが単語群としてうまく出てきている傾向が見られた.

しかし表 1,表 2 の情報どちらにも有益な情報が含まれている中で,前処理だけでは除去しきれなかったノイズや,単語群のずれはゼロにすることができなかった.こういったノイズは主に一つのコメントに対して文脈を考慮していないことが要因として考えられる.さらにサブトピック内の「金-木-雅之」という結果に関してはネット用語特有のノイズといえるような形として単語群が形成された.これは「鈴木さん神」をあえて「金令木さん　ネ申」のように漢字を切り離す書き方をすることによってコメントを大きく見せるというものを判別しきれなかったことによるものである.これらに関しては前処理の段階で特別に処理することでしか改善できないのではないかと考えられる.

実験からは視聴者が全体的に見てどういった内容にいいねしやすいのかに加えて共起語の付加情報によりトピックとして認知しやすい結果となった.しかし結果のノイズの処理に関しては新たに考慮しなければならない課題も出てきたので、解決策を模索していきたい.

4. おわりに

本研究では, YouTube のコメント欄を用いて動画概要の抽出を行った. コメント欄の内容だけでなくいいね数を加味することで概要抽出の幅やトピックの具体性を広げることができたが, ノイズの処理や単語解釈のずれは改善する必要がある. そのため, 今後は文脈の考慮をふまえた実験や前処理にかかわる部分,いいね数の重みづけの面での改善をしたい.

参考文献

[1] David M. Blei, Jon D. McAuliffe. Supervised topic models. In Advances in neural information processing systems (pp. 121-128) 2007.
[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. JMLR, 3:993-1022, 2003.
[3] 鈴木雅之 - 違う,そうじゃない / THE FIRST TAKE https://youtu.be/PmEpx-Ywr2w (2022年12月22日閲覧)