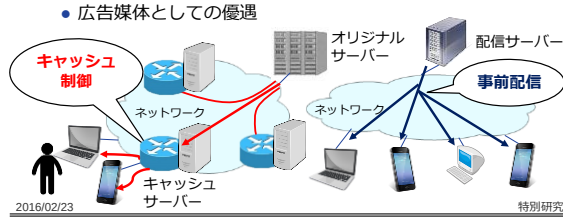
Osaka University

YouTube コンテンツの視聴数推移パターンの分析と人気推移予測手法の提案

大阪大学　基礎工学部　情報科学科
村田研究室　田中 達也

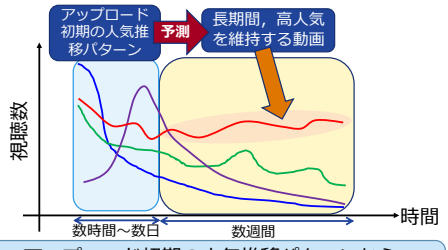
研究背景

- YouTube に代表される UGC (User Generated Content) の視聴の普及
- 長期間高人気が維持される動画をアップロード初期に予測することが望ましい
  - ネットワークトラヒック削減のためのキャッシュ制御
  - 配信サーバのピーク時負荷抑制のための事前配信
  - 広告媒体としての優遇



研究課題

- UGCの将来の人気度予測は困難
  - 多種多様なユーザーが生成するため人気推移パターンが複雑



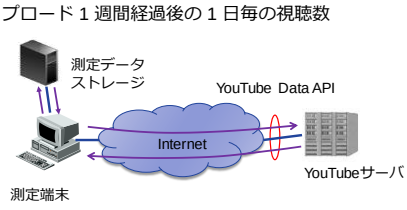
アップロード初期の人気推移パターンから
長期間，高人気を維持する動画を予測できないか？

研究の目的と方法

- 研究目的
  - UGC のアップロード初期の人気推移パターンを分析
  - アップロード初期の時点で将来高人気を維持する動画の予測
- 研究の手順
  - YouTube 動画の視聴数の時系列データの収集
  - **k-means 法**による動画の視聴数推移パターンの傾向分析
    - 視聴数推移パターンに応じたクラスタリング
  - **単純ベイズ分類器**を用いた高人気動画予測の評価
    - 高人気動画の予測を教師あり学習で実現

YouTube データの概要

- 新着動画の視聴数の時系列データ
  - YouTube Data API version3 [13] を用いて取得した動画
    - (動画数：87,830　2015/10/14～2015/12/16)
  - アップロード 1 週間までの 1 時間毎の視聴数
  - アップロード 1 週間経過後の 1 日毎の視聴数

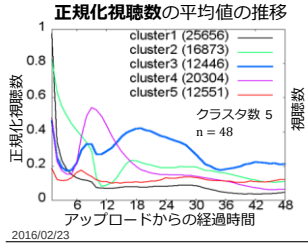


[13] "YouTube Data API" https://developer.google.com/youtube/v3/

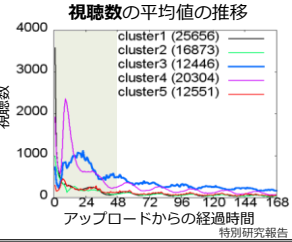
k-means 法を用いた視聴数推移パターンの分析

- アップロード初期の視聴数推移パターンを分類
  - 各動画に対して，最初の n 時間の視聴数の最大値で各時間の視聴数を割った正規化視聴数をもとに k-means 法を適用
- 視聴数を維持する推移パターン (クラスタ 3) の存在を確認

アップロード後 48 時間の
正規化視聴数の平均値の推移

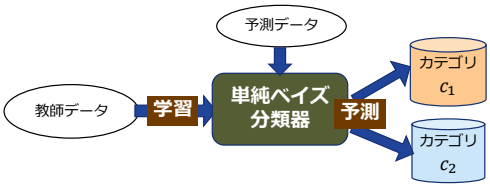


アップロード後 1 週間の
視聴数の平均値の推移



単純ベイズ分類器 (NBC : Naive Bayes Classifier)

- ベイズの定理を用いた**教師あり学習**による分類法
  - **学習**：教師データを用いて各カテゴリ c の学習サンプルがある入力セット $f_1, \cdots, f_n$ を有する確率を計算
  - **予測**：$f_1, \cdots, f_n$ を有する各予測サンプルを事後確率が最大になる c に分類
    - $\text{classify}(f_1, \cdots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$



単純ベイズ分類器の学習と予測の方法

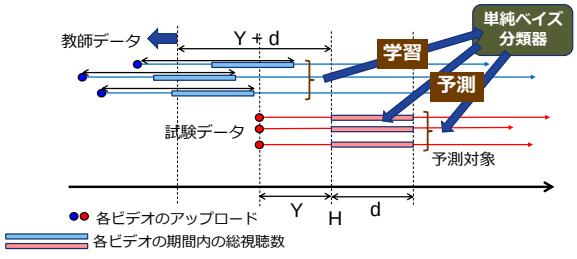
- 入力
  - アップロード初期 Y 時間の視聴数の最大値で各時間の視聴数を割った**正規化視聴数**と**視聴数の最大値の桁数**
- 出力
  - C1：**高人気維持**　C2：それ以外
    - C1 の定義 1 : d 日後までの d 日間の累積視聴数が全体の上位 1%
    - (C1 の定義 2 : d 日後の 1 日の視聴数が全体の上位 1%)

学習データの例

動画 ID	Y 内の正規化視聴数				Y 内の最大視聴数の桁数	分類カテゴリ
	スロット 1	スロット 2	…	スロット Y		
abcdefghijk	1.0	0.5	…	0.4	5	C1
lmnopqrstuv	0.5	0.2	…	0.0	3	C2
wxyz1234567	0.2	0.8	…	0.6	4	C1

単純ベイズ分類器を用いた高人気動画予測

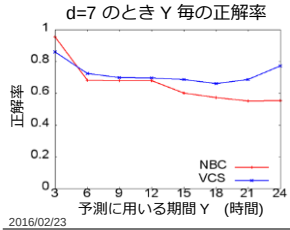
- H：現在時刻，Y：予測に使用する期間，d：予測対象日
- 高人気動画の定義
  - H から d 日間の累積視聴数が全体の上位 1 %



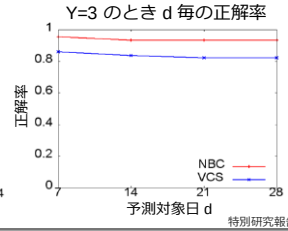
評価結果

- Y = 3 のとき単純ベイズ分類器によって正解率が向上
  - 比較：視聴数上位選択法 (VCS : View Count based Selection)
    - 期間 Y の累積視聴数の多い順に同数の動画を選択
  - スロット長 Y/3 , 学習には常に 3 スロット使用
  - 教師データ、試験データをランダムに半数ずつ選択

d=7 のとき Y 毎の正解率



Y=3 のとき d 毎の正解率



まとめと今後の課題

- まとめ
  - k-means 法を用いた視聴数推移パターンに関する分析
    - 高人気が維持する動画，初期は人気が高いがその後急減する動画が存在することを確認
  - 単純ベイズ分類器を高人気維持動画の予測に適用
    - 初期の人気推移パターンを考慮することで，考慮しないで視聴数の大きいものを単純に選択した場合よりもアップロード初期の時点での予測精度が向上することを確認
- 今後の課題
  - 視聴数以外のメトリクスを利用した人気推移予測
  - 他の教師あり学習手法による人気推移予測
  - キャッシュ制御や事前配信に応用した場合の効果の分析