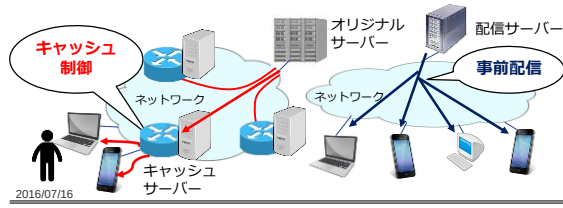
Osaka University

ユーザー生成コンテンツの視聴数推移パターン分析と人気推移予測

田中 達也† 阿多 信吾‡ 村田 正幸†
† 大阪大学 大学院情報科学研究科
‡ 大阪市立大学 大学院工学研究科

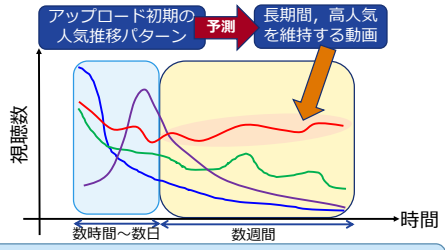
研究背景

- YouTube に代表されるユーザー生成コンテンツ (UGC ; User Generated Content) の視聴の普及
- **人気コンテンツを早期かつ高精度に判別**することが有効
  - ネットワークトラヒック削減のための適切な**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**
  - 効果的な**広告マーケティング**



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



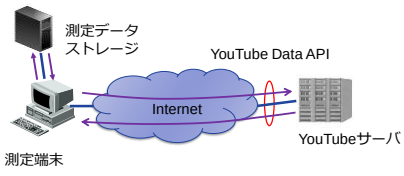
アップロード初期の人気推移パターンから
長期間，高人気を維持する動画を予測できないか？

研究の目的と方法

- 研究目的
  - UGC の人気推移パターンの分析
  - 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 動画の視聴数時系列データの収集
  - 人気度推移の分析
  - クラスタリングによる視聴数推移パターン分析
    - **k-means 法**を用いたクラスタリング
  - 人気度予測と評価
    - 教師あり学習の一種である**単純ベイズ分類器**により判別

YouTube データの概要

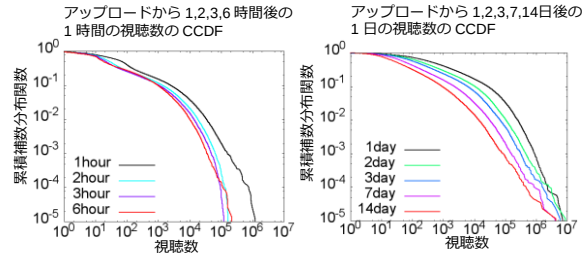
- 新着動画の視聴数の時系列データ
  - YouTube Data API version3 [13] を用いて取得した動画
    - (動画数：87,830　2015/10/14～2015/12/16)
  - アップロード 1 週間までの 1 時間毎の視聴数
  - アップロード 1 週間経過後の 1 日毎の視聴数



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

アップロードから経過時間による視聴数分布

- アップロード直後が最も視聴数が多く、その後減少
- 両対数グラフで裾野の部分が線形に近いカーブ
  - 非常に視聴数の高い少数の動画が存在

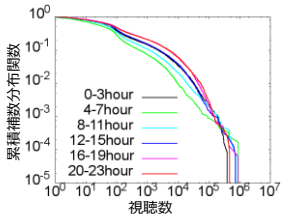


Osaka University　7

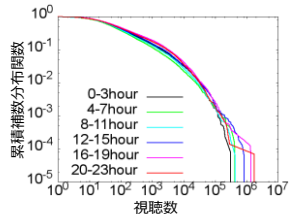
アップロード時刻による視聴傾向の差異

- 世界標準時刻で16-19時、20-23時の視聴数が多い
  - **アップロード直後**はインターネット人口の多い時間帯にアップロードされた動画が多くの視聴数を獲得する傾向

アップロード時間帯別の**アップロード 1 時間後**の視聴数の CCDF



アップロード時間帯別の**アップロード 7 日後**の視聴数の CCDF



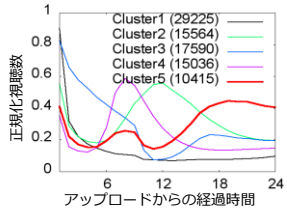
2016/07/16

Osaka University　8

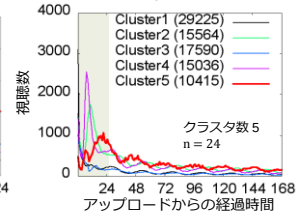
クラスタリングによる視聴数推移パターン分析

- k-means 法を用いて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 5) の存在を確認

アップロード後 24 時間までの**正規化視聴数**の平均値の推移



アップロード後 1 週間までの**1 時間の視聴数**の平均値の推移



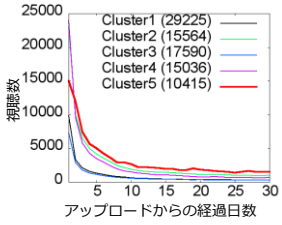
2016/07/16

Osaka University　9

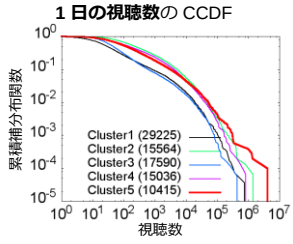
各クラスタの人気度推移

- 初期の時点で正規化視聴数が維持されているクラスタはその後も他のクラスタより視聴数が高い傾向
  - クラスタ 5 が視聴数を維持

1 日の視聴数の平均値の推移



アップロード後 7 日後時点の**1 日の視聴数**の CCDF



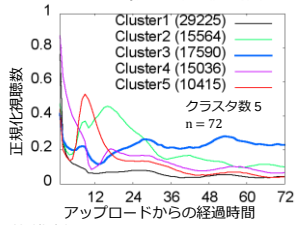
2016/07/16

Osaka University　10

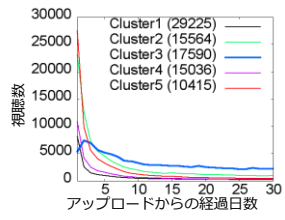
クラスタリングに用いる期間の変更

- n=48,72で k-means 法を用いた場合
  - n=24 の場合と同様の傾向を確認
- n=72 のときクラスタ 3 が**視聴数を維持する推移パターン**

アップロード後 72 時間までの**正規化視聴数**の平均値の推移



1 日の視聴数の平均値の推移

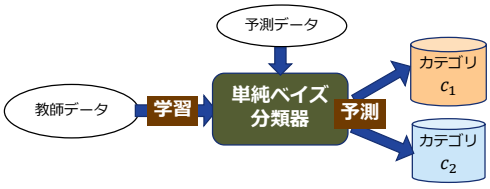


2016/07/16

Osaka University　11

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
    - classify$(f_1,\cdots,f_n) = \arg\max_c p(C=c)\prod_{i=1}^{n} p(F_i = f_i|C=c)$



2016/07/16

Osaka University　12

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

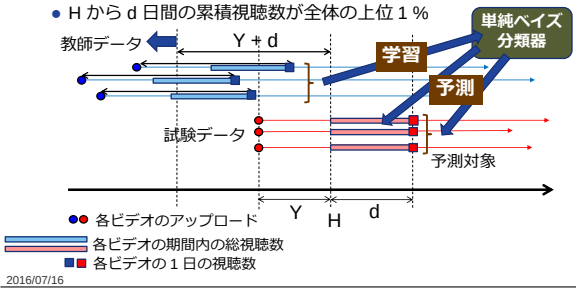
動画 ID	Y 内の正規化視聴数				Y 内の最大視聴数の桁数	分類カテゴリ
	スロット 1	スロット 2	…	スロット Y		
abcdefghijk	1.0	0.5	…	0.4	5	C1
lmnopqrstuv	0.5	0.2	…	0.0	3	C2
wxyz1234567	0.2	0.8	…	0.6	4	C1

2016/07/16

Osaka University　13

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 日後の 1 日の視聴数が全体の上位 1%
  - H から d 日間の累積視聴数が全体の上位 1 %



2016/07/16

Osaka University　14

単純ベイズ分類器による予測結果

- Y = 3 時間のとき単純ベイズ分類器によって予測精度が向上
  - 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
  - 教師データ、試験データをランダムに半数ずつ選択
  - 評価指標：適合率
  - ターゲット
    - 8日目、15日目の 1 日の視聴数が上位 1% の動画
    - 2~8 日目、2~15 日目までの累積視聴数が上位 1% の動画

Y=3 時間のときの人気度予測の結果

	視聴数上位 1% の動画予測の適合率			
ターゲット日	8 日目	2~8 日目	15 日目	2~15 日目
NBC	0.785	0.956	0.707	0.933
VCS	0.697	0.860	0.674	0.837

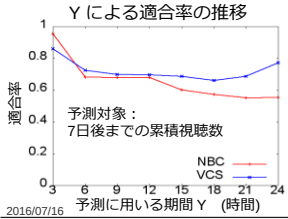
2016/07/16

Osaka University　15

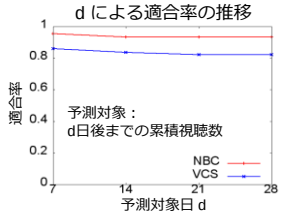
評価結果

- 予測対象日を d=7 と固定し、Y を変更した場合
  - Y が大きくなるにつれ NBC の適合率が減少
    - スロット長 Y/3 , 学習には常に 3 スロット使用
- 予測に用いる期間を Y=3 と固定し、d を変更した場合
  - d の値にかかわらず、NBC が優勢

d=7 のとき
Y による適合率の推移



Y=3 のとき
d による適合率の推移



2016/07/16

Osaka University　16

まとめと今後の課題

- まとめ
  - 視聴数分布に関する分析
    - アップロード直後の視聴数が多く、時間経過につれ減少
    - ごく少数の動画が非常に多い視聴数を獲得
  - 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 3 時間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
  - 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

2016/07/16