特別研究報告

題目

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

指導教員
村田 正幸 教授

報告者
田中 達也

2016 年 2 月 16 日

大阪大学 基礎工学部 情報科学科

平成 27 年度 特別研究報告

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

田中 達也

内容梗概

近年，YouTube 動画に代表される UGC（User Generated Content）の視聴がインターネット上で人気のあるサービスとなってきている．メディア広告の配置やトラヒックコントロール，コンテンツキャッシュなど多くの面から，コンテンツの将来の人気度の予測が有用である．動画配信において CDN（Content Delivery Networks）を運用する形態も増えてきているが，ネットワーク内のトラヒックを効果的に削減するという観点からキャッシュ制御を適切に行う必要がある．従来のキャッシュ制御法である LRU（Least Recently Used）はコンテンツの将来の人気度を考慮しないで直近の人気度が高いコンテンツを優先的にキャッシュするが，キャッシュの利用効率を高めるためには将来の人気度を考慮してキャッシュ制御を行うことが望ましい．また，動画配信のトラヒックによる配信サーバやネットワークのピーク負荷を低減するためには，高人気の動画をユーザ端末に事前配信することも有効であるが，事前配信時の負荷を抑えるためには，できるだけ高い人気が継続することが予想される動画のみに限定して事前配信を行う必要がある．また，UGC の投稿者は広告の掲示が可能であり，視聴数や広告の種類に応じて報酬を得ることが可能であるが，YouTube などのような CGM（Consumer Generated Media）にも広告主から料金が支払われていると推測できる．視聴数が多い動画，つまり人気度が高い動画は CGM にとって広告収入を多く得ることができる優良顧客である．CGM にとって，人気度が高く，視聴数が長期間継続して多く得られる動画の投稿者を優遇することで，高い広告収入が見込まれるコンテンツを多く集めることが可能となる．

これらの観点から，できるだけアップロード初期の段階で，コンテンツの将来の人気度の変化パターンを予測して，将来も高い人気が維持される動画を予測することが有効であるが，UGC の視聴数推移は，莫大な動画数，質・内容の多様性といった理由から将来の予測が難しい．

本報告では UGC として代表的な YouTube の各動画の視聴数を実測し，視聴数の推移パターンに基づき k-means 法を用いて YouTube 動画をクラスター分析することで，YouTube

の各動画の人気の推移パターンに関する傾向を明らかにする．また，アップロード初期の視聴数の推移パターンと視聴数の絶対値から，長期間にわたり高人気を維持する動画を予測する手法として，教師あり機械学習法の一種である単純ベイズ分類器を適用した場合の予測精度を評価する．その結果，アップロード初期 3 時間の視聴数データを用いて予測するとき，単純ベイズ分類器で初期の変化パターンを考慮することで変化パターンを考慮しないで視聴数の絶対値のみを予測する場合と比較して，正解率が約 10 %向上することを明らかにした．

主な用語

YouTube
人気度予測
視聴数推移パターン
k-means 法
単純ベイズ分類器

目 次

図 目 次

表 目 次

1　はじめに

近年，YouTube[1] に代表される UGC（User Generated Content）の視聴がインターネット上で人気のあるサービスとなってきている．動画ストリーミング配信は，遅延時間に対する要件が厳しく，一定のタイミングでデータが届く必要があり，ネットワークの輻輳などの要因によって遅延が増加した場合に，ユーザが体感する視聴品質の劣化度合いが大きく，安定して良好な視聴品質を維持することが重要である．

遅延軽減のため，OTT（Over The Top）事業者 (YouTube など) は，ネットワーク上の複数の箇所に分散配置したキャッシュサーバを用いてコンテンツを配信しているが，ユーザの安定した視聴品質を維持するためには，キャッシュの効果が高いと思われるコンテンツを優先してキャッシュするなど，適切な制御が必要である．また近年，ISP（Internet Service Provider）が CDN（Contents Delivery Networks）を運用する形態も増えているが，動画ストリーミング配信のトラヒック量は大きいことからネットワークに与えるインパクトが大きいため，ネットワーク内のトラヒックを効果的に削減するという観点からもキャッシュ制御を適切に行う必要がある．しかし従来のキャッシュ制御法は，LRU（Least Recently Used）等のシンプルなものが使用されており，過去のアクセスパターンに基づきキャッシュを制御しているが，本来は，将来の人気度を考慮してキャッシュ制御を行うことが望ましい．

また，動画配信サーバのピーク時負荷を抑制するためには，低需要時に高人気と予測される動画を事前に配信することが有効であるが，事前配信量を抑えつつ高いピーク時負荷抑制効果を得るためには，人気が継続することが予想される動画のみを限定して配信する必要がある [2]．

UGC の投稿者は広告の掲示が可能であり，視聴数や広告の種類に応じて報酬を得ることが可能であるが，CGM（Consumer Generated Media) 側にも広告主から料金が支払われていると推測できる．高い視聴数が長期間継続することを見込める動画は，CGM にとって優良動画であり，投稿初期の時点で注目動画のリストに掲載するなどの優遇を行うことが望ましい．

そのため，長期間にわたり高い人気が維持される動画をアップロード初期に予測できることが望ましい．UGC の視聴数推移は，コンテンツプロバイダによって商用サービスとして提供される動画視聴サービスである VoD（Video-on-Demand）に比べ，莫大な動画数，質・内容の多様性といった理由から，将来の人気度の予測が難しく，また動画ごとに人気が大きく異なることが知られており，様々な文献で UGC の視聴傾向の分析が行われている．

文献 [3] はソーシャルニュースサイト Digg[4] において新しく投稿された記事に対するユーザーの初期の反応から，ユーザーインターフェースの特徴をモデルとして組み込んで，記事が人気になるかどうかを予測している．文献 [5] は YouTube の動画を人気ランキングに属す

る動画，著作権侵害により削除された動画，YouTube の検索エンジンにランダムな語彙を入力することにより選ばれた動画の 3 つのデータセットに分け，それぞれについて視聴数の推移を調査している．文献 [6] は YouTube からランダムサンプリングしたコンテンツの，1 週間の粒度で見たアクセス数の推移を分析し，各動画の視聴数が最大となった週，それより前の週，それより後の週でアクセス数の分布が異なることを示し，その知見を踏まえてモデル化している．文献 [7] は YouTube のアクセスパターンを分析し、多くのコンテンツの日々の視聴数の変化パターンは，大きく長期間アクセスと単発的の二つのタイプに分類できることを示し，長期間アクセスのタイプについて，主成分分析を用いて将来の視聴数を予測する方法を提案している．文献 [8] は Digg と YouTube のアクセスパターンを分析し，動画アップロード後，早期の視聴数と 30 日後の視聴数が対数グラフにおいて線形の相関を持つことに着目し，テストセットで線形モデルのパラメータ調整を行うことによって，初期の視聴数から将来の視聴数を予測できることを示している．

文献 [9] は文献 [8] の方式における，アップロード x 日時点の累積視聴数が同じであっても，予測ターゲット日における累積視聴数が大きく異なる動画がある問題に着目し，アップロードから x 日までの人気度の推移パターンから，任意のターゲット日までの累積視聴数を線形回帰モデルで予測し，その予測が文献 [8] の予測よりも優れていることを示している．文献 [10] は YouTube の基本的な人気度を表す視聴数と，コメント数，お気に入り数，評価数といった人気度を表す別のメトリクスとの間には相関があることを明らかにし，コメント数，お気に入り数，評価数から視聴数を推定する単純な線形回帰モデルを提案している．文献 [11] はトラフィック量の予測という観点から，コンテンツを局所性に応じてグループ化し，自己回帰モデルを用いたマルチモデルで短期的なアクセスパターンの予測を行っている．文献 [12] では，初期の 1 日の粒度でみた視聴数の変化パターンを k-means 法を用いて分類し，高い視聴数が長期間維持される動画の集合を抽出することを提案している．しかしこれら既存の方式では．小さい計算量で高人気が持続する動画を高精度に予測することは困難である．

本報告では，まず，YouTube 動画の視聴数の時系列データを収集し，視聴数の推移パターンの傾向を分析する．文献 [12] では 1 日の粒度の視聴数の変化パターンを分析しているが，本報告では，1 時間単位という微細化した時間粒度で時系列データを収集し，k-means 法を用いた動画の分類を行い，細かい時間粒度での視聴数変化パターンの傾向を明らかにする．また，初期の一時間毎の視聴数の推移パターンと視聴数の絶対値から，以後長期間にわたって高人気を維持する動画を予測する方法として，教師あり学習の一種である単純ベイズ分類器を適用した場合の予測精度を評価する．

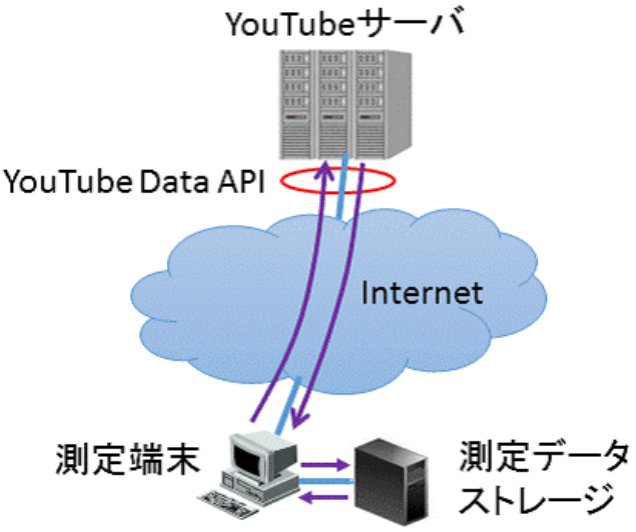


図 1: データ測定方法

2　YouTube 視聴数の測定と分析

本章では，分析に使用する YouTube 視聴数データセットの概要と測定方法について述べる．次に YouTube 動画の視聴数の時系列データを用いて，YouTube の人気度や視聴数推移パターンの傾向について分析する．

2.1　視聴数データ測定方法

YouTube 動画の人気度と視聴数の推移パターンの傾向分析，単純ベイズ分類器の評価のために，次の YouTube 動画データセットについて，図 1 に示すように YouTube が提供している API（YouTube Data API version3.0）[13] を用いて，アップロードから 1 週間経過までの 1 時間ごとの視聴数，アップロードから 1 週間経過した動画の日々の視聴数の測定を行った．

- **新着動画** - 動画がアップロードされてからの視聴数の推移を調査するためのデータセット．データの取得期間は 2015 年 10 月 14 日から 2015 年 12 月 16 日で，1 か月間動画が存在して，視聴数が継続して取得できた動画である．動画数は 87,830 個である．
- **アップロード国ごとの動画** - アップロードされた国ごとの視聴傾向を調査するためのデータセット．データの取得期間は 2015 年 12 月 25 日から 2016 年 1 月 21 日までで，視聴数が 1 週間継続して取得できた動画である．動画数は 40,900 個である．

新着動画データセットは動画がアップロードされてからの視聴情報の推移パターンを観測するためのデータセットである．このデータセットを構成するために，新たにアップロードされた動画を収集し続けるプログラム，新たにアップロードされた動画がアップロードから 1 週間経過するまで 1 時間毎に視聴数を取得するプログラム，アップロードから 1 週間経過した動画の視聴数を日毎に取得するプログラムを実行した．

以下は，視聴数データ収集プログラムの流れである．

- 1 分ごとに，新着動画 ID を YouTube API に問い合わせて取得し，取得動画 ID を，動画 ID リストに追記 (動画 ID リストは分ごとに用意)
- 各動画について，アップロード後 1 週間以内は，1 時間ごとに，それまでの累積視聴数を YouTube API に問い合わせ，取得累積視聴数データを，視聴数時系列データファイルに追記
- 各動画について，アップロード後 1 週間以後は，1 日ごとに，それまでの累積視聴数を YouTube API に問い合わせ，取得累積視聴数データを，視聴数時系列データファイルに追記
- 各動画について，各時点の累積視聴数から，直前の値を引くことで，各 1 時間もしくは一日の視聴数データを生成

このプログラムで取得した視聴数の時系列データのうち，以下の条件を満たす 87,830 個の動画をデータセットとして使用する．

- アップロードから 1 週間経過まで 1 時間間隔で視聴数を取得できた動画
- アップロードから 1 週間経過後は 1 日間隔で 30 日まで視聴数を取得できた動画
- アップロード 8 日経過時での累積視聴数が 10 以上
- 1 日あたりの視聴数が 5 以下の日が 7 日間連続して続かない
- 1 時間間隔の視聴数がマイナスにならない

YouTube では，動画が不正な方法で視聴数を得ていないかを常にチェックしており，不正な方法であると判断した場合，動画の視聴数の修正を行う場合がある [14]．そのため累積視聴数が減少し，差分をとり 1 時間間隔の視聴数をみたときにマイナスになっている場合がある．このような動画は分析対象から外している．

アップロード国ごとの動画データセットはアップロード国別に視聴数時系列データである．このデータセットを構成するために，指定した経度緯度地点から，指定した半径 (最大

1,000km) 以内の場所で，現在時刻の 30 分前にアップロードされた動画の ID を毎分取得するプログラムを作成した．取得した動画 ID を**新着動画**データセットの視聴数時系列データ取得方法と同様に，1 週間分の視聴数時系列データを取得したものを**アップロード国ごとの動画**データセットとして用意する．以下の地域に対して，各国のエリアが入るような中心座標と半径を設定し，取得した．

- 日本（**JP**）
- アメリカ西部（**WestUS**）
- アメリカ東部（**EastUS**）
- イギリス（**GB**）
- フランス（**FR**）
- ドイツ（**DE**）
- ブラジル（**BR**）
- インド（**IN**）

2.2 視聴数分布の時間推移に関する分析

本節では，視聴数分布の時間推移に関する分析を行う．まずアップロードから一定期間後（1 時間，2 時間，3 時間，6 時間）の 1 時間の視聴数の累積補分布を図 2，アップロードから一定期間後（1 日，2 日，3 日，7 日，14 日）の 1 日の視聴数の累積補分布を図 3 に示す．どちらのグラフも両対数グラフで，裾野の部分が線形に近いカーブとなっている．これは極端に視聴数が多い，少数の動画が存在していることを示している．また図 2 より，アップロード直後の 1 時間の視聴数が最も大きい．これはアップロード直後は SNS（Social Networking Service）上などで，動画に関する情報が拡散され多くの人に視聴されるためであると予想される．また図 3 よりアップロードから時間が経過するにつれて視聴数が減少していくこと，視聴数の大きい動画がより少数になっていくことがわかる．アップロードから一定期間（1 日，2 日，3 日，7 日，14 日）の累積視聴数の累積補分布を図 4 に示す．この図も値の大きい領域で線形に近いカーブ有しており，極端に視聴数の多い，少数の動画が存在していることがわかる．

次に動画がアップロードされた時間帯別（世界標準時で 0〜3 時，4〜7 時，8〜11 時，12〜15 時，16〜19 時，20〜23 時）に，アップロードから 1 時間後の視聴数の累積補分布を図 5，

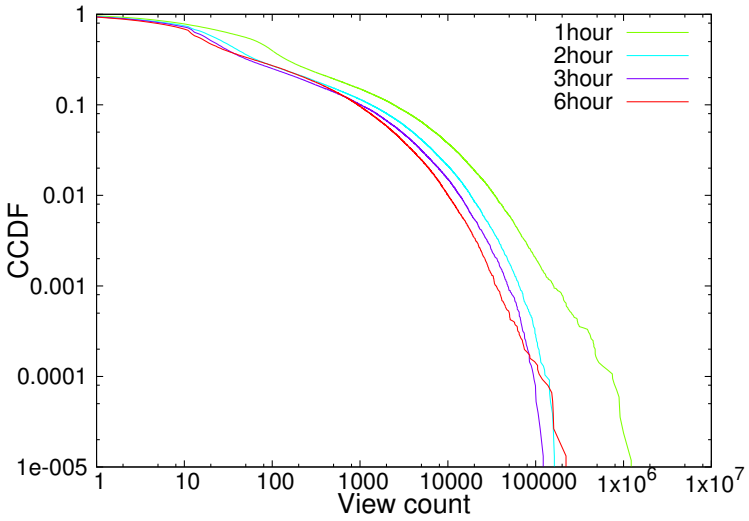


図 2: アップロードから 1，2，3，6 時間後の 1 時間の視聴数の CCDF

アップロードから 1 日後の視聴数の累積補分布を図 6，アップロードから 7 日後の視聴数の累積補分布を図 7 に示す．これらの図から，アップロード直後はアップロード時間帯による差が大きく，時間が経過するにつれて差が小さくなっていくことが確認できる．アップロード直後 1 時間の間は，時間帯によりインターネットの利用者数が異なるため視聴数に差が表れるが，時間が経過するにつれて，アップロード時刻にかかわらず，コンテンツそのものの人気度により視聴数が決まると考えられる．比較的視聴数の多い時間帯は世界標準時で 20〜23 時と 16〜19 時である．これはヨーロッパ圏の夕方から夜の時間帯になりインターネット利用が多い時間帯である．このことからヨーロッパ圏でアップロードされた動画の視聴数が多くなる傾向があると予想される．アップロードから 7 日間の累積視聴数の累積補分布を図 8 に示す．図 7 と同様の傾向が確認できる．

次に国別の累積補分布を調査する．アップロードから 1 時間後の視聴数の累積補分布を図 9，アップロードから 1 日後の視聴数の累積補分布を図 10，アップロードから 7 日後の視聴数の累積補分布を図 11，アップロードから 7 日間の累積視聴数の累積補分布を図 12 に示す．図中の凡例は国名とその国の取得できた動画数を示す．これらの図から，どの国でも視聴数の増加に対して累積補分布は緩やかに減少しており，極端に視聴数が多い，少数の動画が存在していることがわかる．また，アメリカなどの英語圏でアップロードされた動画は，同一言語を使用する人口が多いため全体的に視聴数が多い反面，日本など，同一言語を使用する人口が少ない国でアップロードされた動画は視聴数が少ない傾向にあることがわかる．

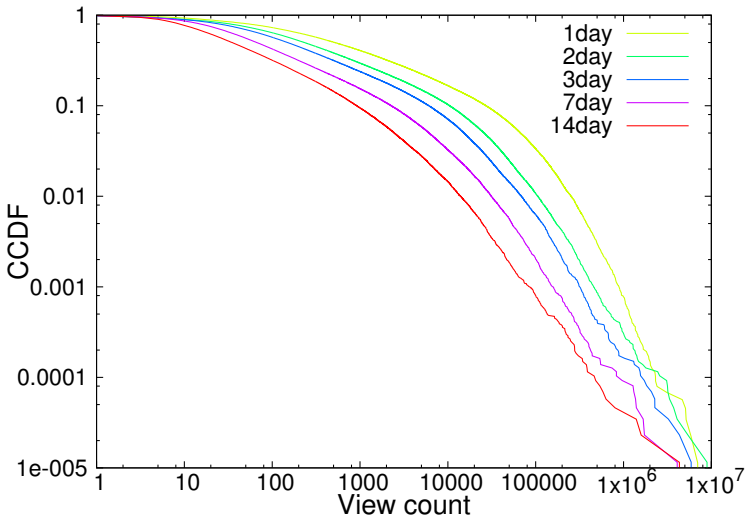


図 3: アップロードから 1，2，3，7，14 日後の 1 日の視聴数の CCDF

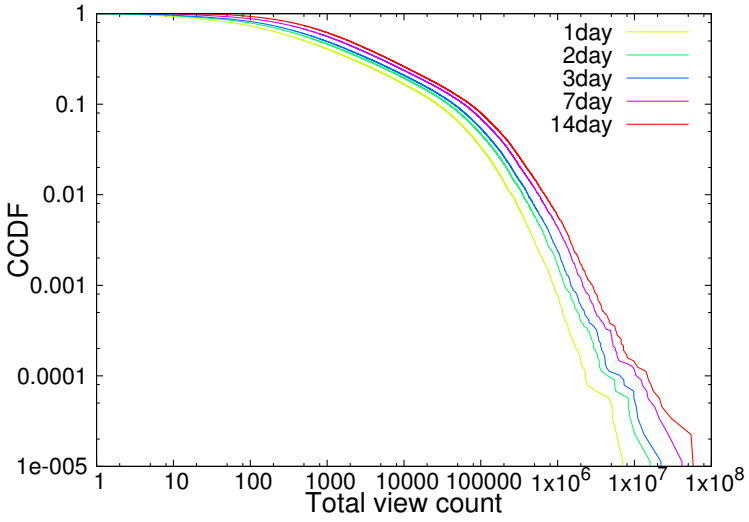


図 4: アップロードから 1，2，3，7，14 日間の累積視聴数の CCDF

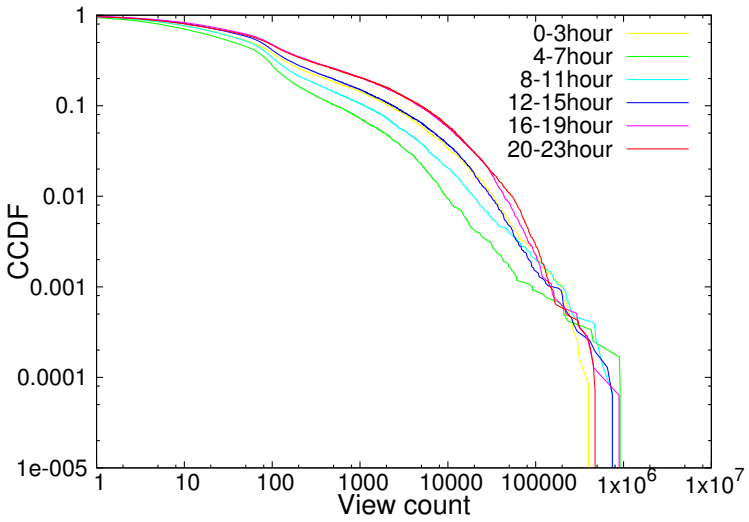


図 5: アップロード時間帯別のアップロードから 1 時間後の視聴数の CCDF

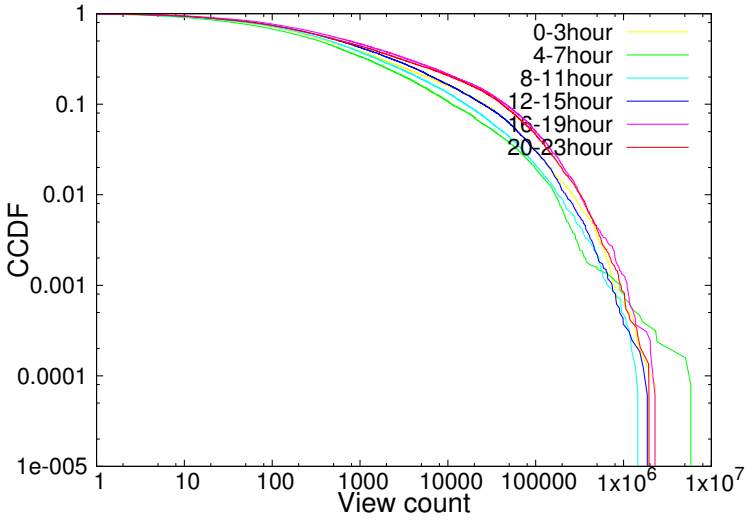


図 6: アップロード時間帯別のアップロードから 1 日後の視聴数の CCDF

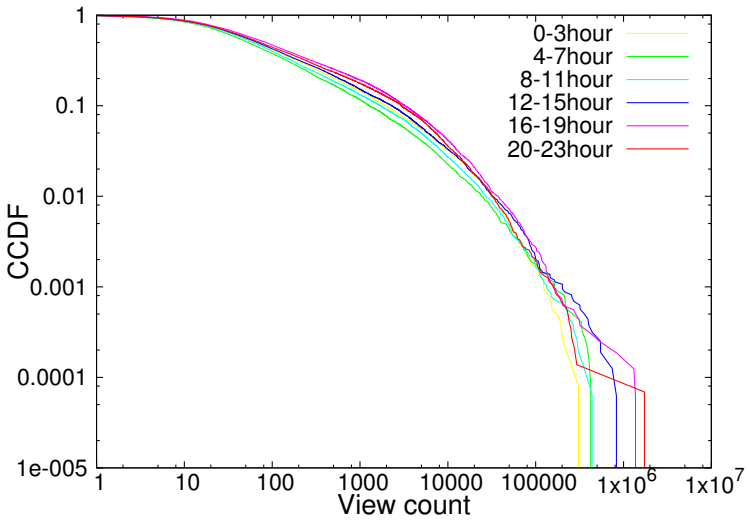


図 7: アップロード時間帯別のアップロードから 7 日後の日視聴数の CCDF

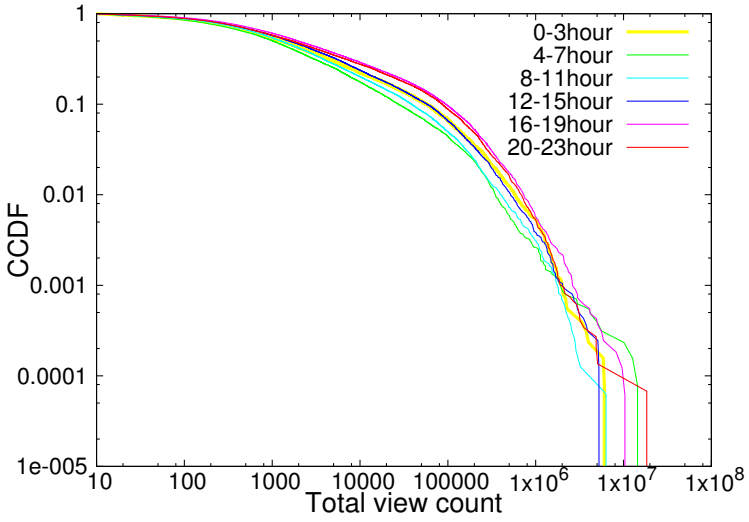


図 8: アップロード時間帯別のアップロードから 7 日間の累積視聴数の CCDF

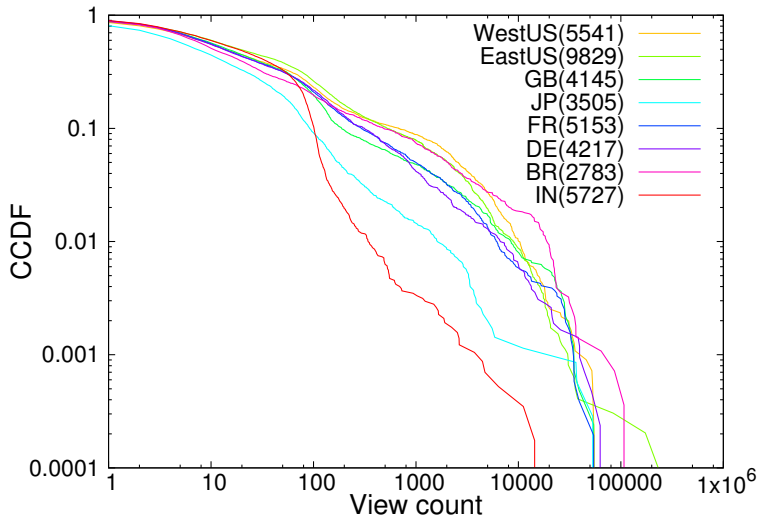


図 9: アップロード国別のアップロードから 1 時間後の視聴数の CCDF

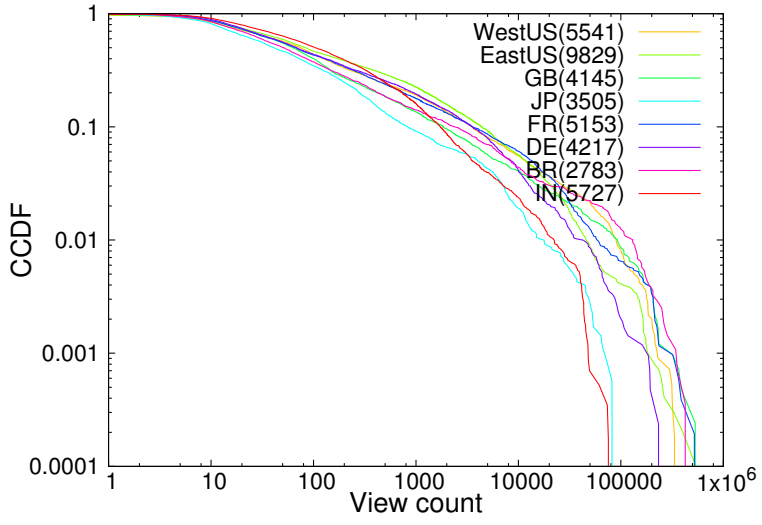


図 10: アップロード国別のアップロードから 1 日後の視聴数の CCDF

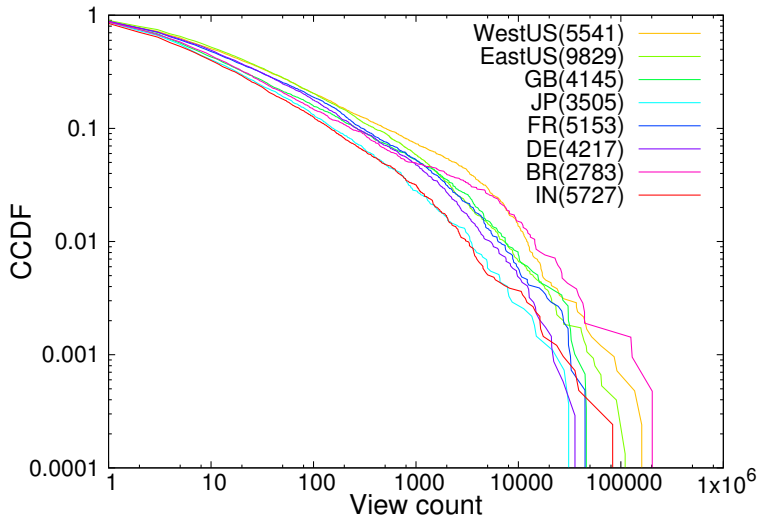


図 11: アップロード国別のアップロードから 7 日後の日視聴数の CCDF

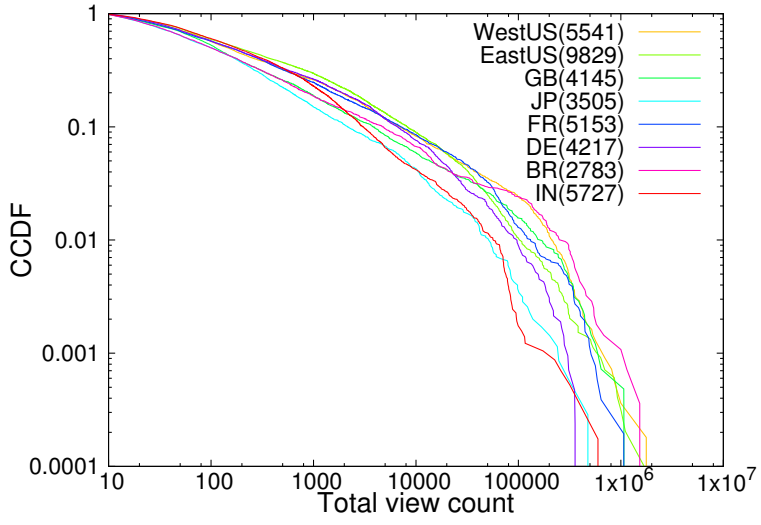


図 12: アップロード国別のアップロードから 7 日間の累積視聴数の CCDF

2.3 k-means 法を用いた視聴数推移パターンに関する分析

アップロードされてから最初の n 時間における各時間の視聴数に対し，非階層型クラスタリング法の代表的な手法である k-means 法を用いて，YouTube の各動画の人気度の推移パターンに関する傾向を分析する．

文献 [12] では 1 日の粒度の視聴数の変化パターンを k-means 法で分析しているが，本報告では，1 時間単位という微細化した時間粒度で時系列データを収集し，k-means 法を用いた動画の分類を行い，細かい時間粒度での視聴数変化パターンの傾向を明らかにする．具体的には，各動画（総数を v とする）について最初の n 時間の視聴数の最大値で各時間の視聴数を割る．その結果，0 から 1 の範囲の値を要素に持つ n 次元のベクトルが得られる．得られた v 個のベクトルを用いて，k-means 法で各動画をクラスタリングすることで，変化パターンが似ている動画を分類する．

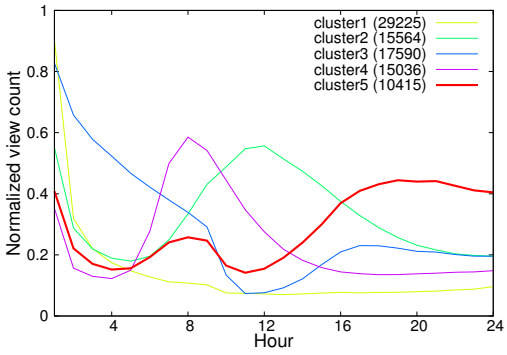
クラスタ数 5 で，アップロードから 24 時間までの視聴数データを用いてクラスタリングを行った結果を図 13 に示す．図の凡例のクラスタ名の横の括弧内の数字は，そのクラスタに分類された動画数である．各クラスタのアップロードからの経過時各時間の正規化視聴数の平均値を図 13(a) に示す．動画数が最も多いクラスタ 1 はアップロード直後の正規化視聴数が高く，それ以降は常に低い値である．つまりアップロード直後のみ視聴され，それ以降は視聴数が維持されない動画が多数を占めていることがわかる．クラスタ 5 はアップロード直後は正規化視聴数は高くないが，16 時間後以降は他のクラスタより正規化視聴数の値が高い．各クラスタのアップロードから 1 時間間隔の平均視聴数の推移を図 13(b) に示す．クラスタ 5 が他のクラスタより高い平均値を維持している．アップロード 24 時間後はおよそ 24 時間周期で視聴数が増減を繰り返していることがわかる．これは，アップロード初期は SNS 上などで動画の存在を知ったユーザーの視聴が多いため，アップロードからの経過時間が視聴数の支配要因であるが，アップロードから 24 時間経過以降は，ユーザーの生活サイクルが支配要因となり，インターネット人口の多い時間帯が視聴数が大きくなるためと考えられる．アップロードから 30 日経過までの日毎の平均視聴数を図 13(c) に示す．図 13(a) において後半の正規化視聴数の値が他のクラスタより高かったクラスタ 5 は他のクラスタより高い視聴数を維持している．アップロード 7 日後の日視聴数の累積補分布を図 13(d) に示す．クラスタ 5 が他のクラスタより高い傾向があり，これらからアップロード初期の 24 時間の間，正規化視聴数が維持されている変化パターンをもつ動画がその後，長期間にわたって高い視聴数を維持する傾向があることがわかる．

次にクラスタ数 5 で，アップロードから 48 時間までの視聴数データを用いてクラスタリングを行った結果を図 14 に示す．各クラスタのアップロードからの経過各時間の正規化視聴数の平均値を図 14(a) に示す．クラスタ 3 は正規化視聴数の減少が他のクラスタより小さ

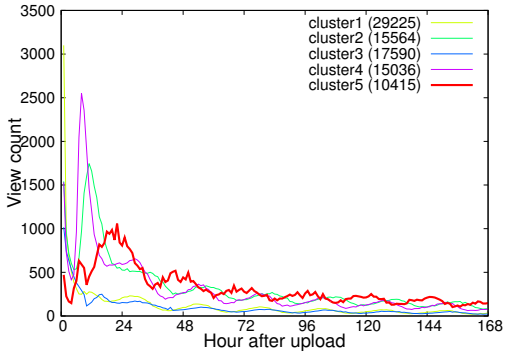
く，高い値を維持していることがわかる．各クラスタのアップロードから 1 時間間隔の平均視聴数の推移を図 14(b) に示す．クラスタ 3 が他のクラスタより高い平均値を維持している．クラスタ 4 はアップロード 12 時間まではクラスタ 3 より高い視聴数であるが，それ以降は減少し，クラスタ 3 より小さくなっている．アップロードから 30 日経過までの日毎の平均視聴数を図 14(c) に示す．クラスタ 3 は他のクラスタより高い視聴数を維持している．アップロード 7 日後の日視聴数の累積補分布を図 14(d) に示す．クラスタ 3 が他のクラスタより高い傾向があり，これらからアップロード初期の間で正規化視聴数が維持されている変化パターンをもつ動画が将来にわたって高い視聴数を維持する傾向があることがわかる．

最後にクラスタ数 5 で，アップロードから 72 時間までの視聴数データを用いてクラスタリングを行った結果を図 15 に示す．各クラスタのアップロードからの経過各時間の正規化視聴数の平均値を図 15(a) に示す．クラスタ 3 はアップロード直後は正規化視聴数は高くないが，後半にかけて他のクラスタより正規化視聴数の値が高い．各クラスタのアップロードから 1 時間間隔の平均視聴数の推移を図 15(b) に示す．後半の正規化視聴数が高かったクラスタ 3 が他のクラスタより平均値の減少がなく維持されていることがわかる．アップロードから 30 日経過までの日毎の平均視聴数を図 15(c) に示す．クラスタ 3 は平均視聴数の減少が少なく視聴数を長期間にわたって維持していることがわかる．アップロード 7 日後の日視聴数の累積補分布を図 15(d) に示す．クラスタ 3 が他のクラスタより高い傾向があり，これらからアップロード初期の 72 時間の間，正規化視聴数が維持されている変化パターンをもつ動画がその後，将来にわたってまで高い視聴数を維持する傾向があることがわかる．しかし，アップロード初期 24 時間や 48 時間でクラスタリングを行ったときと比べ，視聴数が維持される傾向にあるクラスタ 3 が，7 日後の視聴数が $10,000$ 以下の割合が他のクラスタより多くなっている．これはクラスタリングを行う初期の期間が長くなり変化パターン種類が多くなるため，一つのクラスタにさまざまな変化パターンが含まれているためであると考えられる．

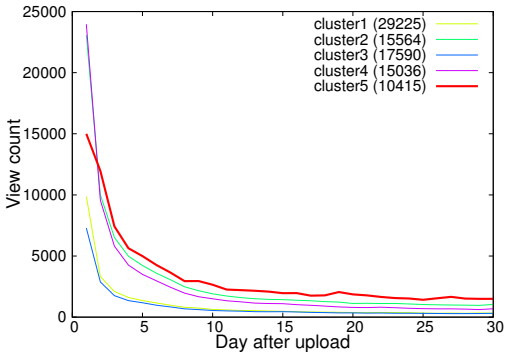
以上のことからアップロード初期の視聴数の絶対値は大きいが長期間維持されない変化パターン，初期の正規化視聴数の減少が小さく将来にわたって人気が維持される変化パターンが存在することが明らかになった．これらを踏まえて，3 章で，将来にわたって高人気が維持される動画を予測する手法として，単純ベイズ分類器を適用した場合の予測精度を評価する．



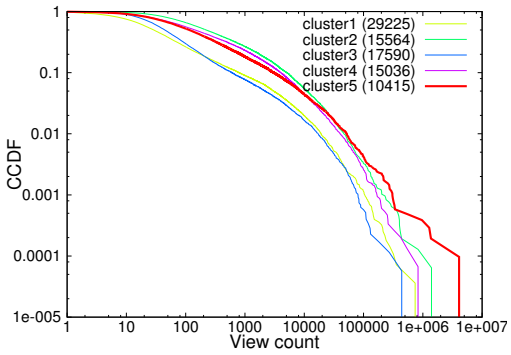
(a) 正規化視聴数の平均値



(b) 1 時間間隔の視聴数の平均値

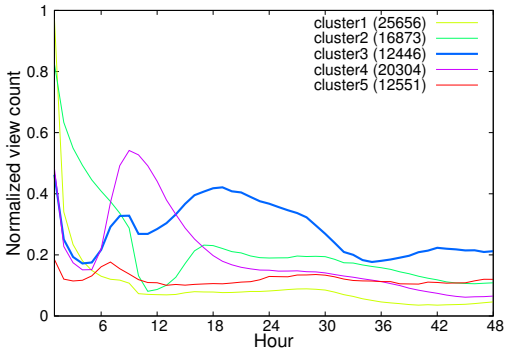


(c) 1 日間隔の視聴数の平均値

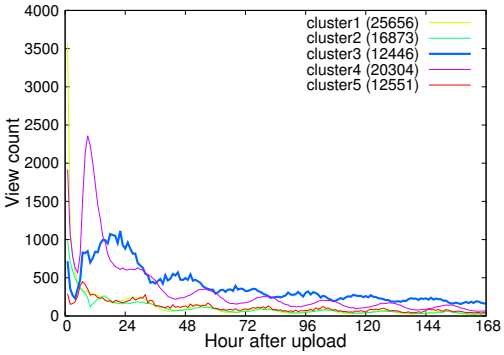


(d) アップロード 7 日後の視聴数の累積補分布

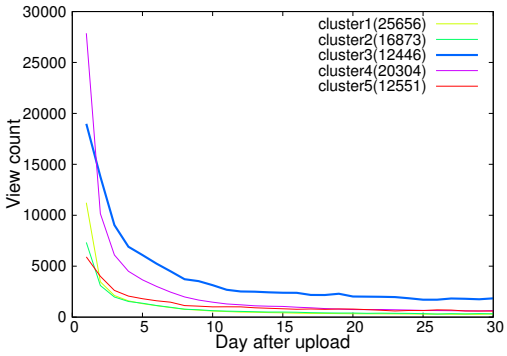
図 13: アップロードから 24 時間までの 1 時間毎の視聴数データに対してクラスタ数 5 で k-means 法を用いたクラスタリングの結果



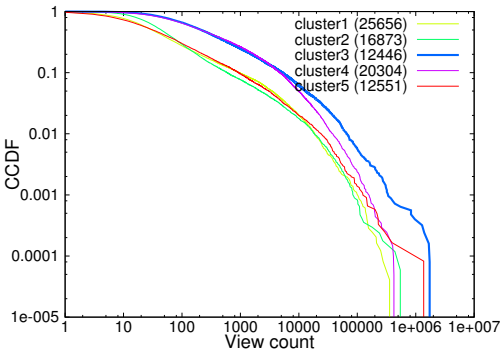
(a) 正規化視聴数の平均値



(b) 1 時間間隔の視聴数の平均値

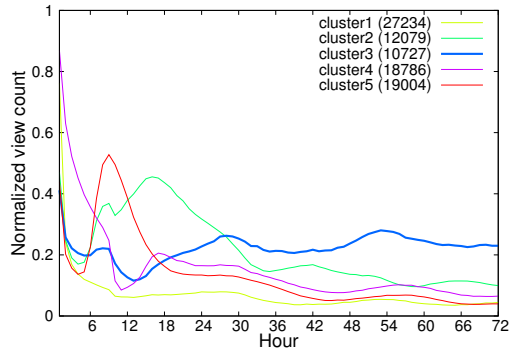


(c) 1 日間隔の視聴数の平均値

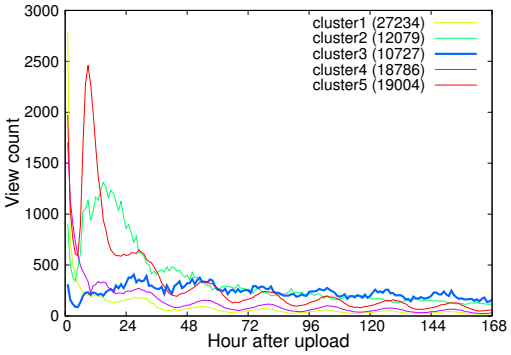


(d) アップロード 7 日後の視聴数の累積補分布

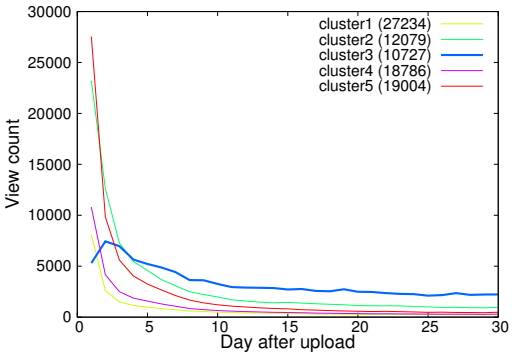
図 14: アップロードから 48 時間までの 1 時間毎の視聴数データに対してクラスタ数 5 で k-means 法を用いたクラスタリングの結果



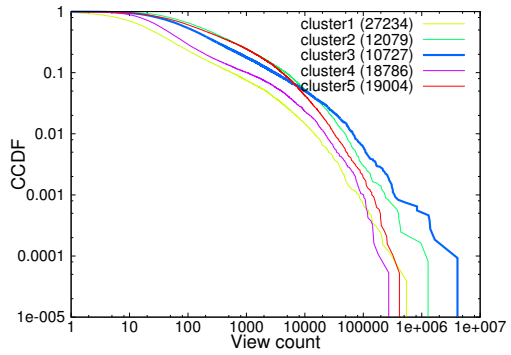
(a) 正規化視聴数の平均値



(b) 1 時間間隔の視聴数の平均値



(c) 1 日間隔の視聴数の平均値



(d) アップロード 7 日後の視聴数の累積補分布

図 15: アップロードから 72 時間までの 1 時間毎の視聴数データに対してクラスタ数 5 で k-means 法を用いたクラスタリングの結果

3　単純ベイズ分類器を用いた高人気 YouTube 動画の予測

本章では，アップロードされてから最初の Y 時間における各時間の視聴数のパターンに対し，将来も視聴数が維持されると予想される動画を予測する方法として，教師あり機械学習の一種である単純ベイズ分類器を用いた場合の評価を行う．

3.1　単純ベイズ分類器の概要

単純ベイズ分類器は，ベイズの定理を適用する教師あり学習の一種である．ベイズの定理は条件付き確率について成り立つ定理で，入力 A が与えられたとき出力 B が得られる確率 $P(B|A)$ を以下の等式

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

で表せる．入力として n 次元の変数 $F_1, \cdots, F_n$ が与えられた時のカテゴリ C に分類される確率 $p(C|F_1, \cdots, F_n)$ はベイズの定理より，

$$p(C|F_1, \cdots, F_n) = \frac{p(C)p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)}$$

と表す．分母は C に依存していないので，ここでは分母を考慮しない．条件付き確率の定理より分子は

$$p(C, F_1, \cdots, F_n) = p(C)p(F_1|C)p(F_2|C)p(F_3|C)\cdots$$

とおける．単純ベイズ分類器は上記の確率モデルにカテゴリの決定規則を合わせたものであり，カテゴリの決定は，事後確率が最大となる C を選択する．この分類器を関数 classify とすると，

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

と表される．

3.2　高人気動画予測のフレームワーク

単純ベイズ分類器を用いた高人気動画予測の全体の流れを図 16 に示す．予測を行う時刻を H，予測に使用するアップロード初期の時間を Y，予測対象日を d 日後とする．H から $Y + d$ 以上前にアップロードされた動画を教師データとして単純ベイズ分類器を学習する．H から Y だけ前にアップロードされた動画を予測対象とし，期間 Y の正規化視聴数と Y 内の最大視聴数の桁数から単純ベイズ分類器で予測を行い，高人気の定義を満たすかどうかを予測する．以下に単純ベイズ分類器を用いた高人気動画予測の流れを記述する．

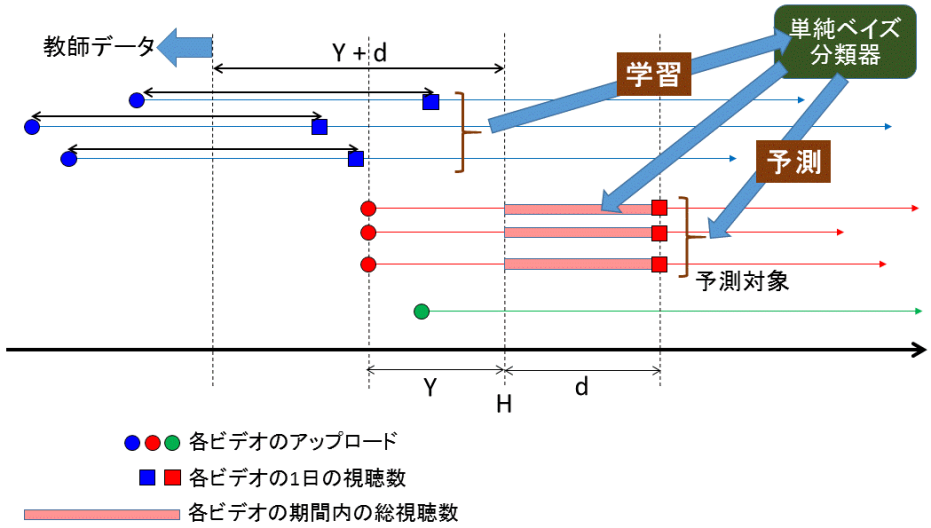


図 16: 単純ベイズ分類器を用いた高人気動画予測の処理サイクル

1. 高人気動画の定義を設定する．
2. YouTube Data API に 1 分周期でアクセスして，新着動画リストを取得する．
3. アップロードから Y 以内の各動画に対して，YouTube Data API に 1 時間周期でアクセスして，1 時間毎の視聴数を取得する．
4. アップロードから Y が経過した動画で経過日数が d 日未満の各動画に対して，YouTube Data API に 1 日周期でアクセスして 1 日ごとの視聴数を取得する．
5. 1 時間周期で，アップロード後から $Y + d$ 以上の時間が経過した動画の初期 Y 期間の視聴数データを用いて，単純ベイズ分類器の学習データを生成する．学習データの例を表 1 に示す．
6. 1 時間周期で，アップロードから Y が経過した動画を対象に，単純ベイズ分類器を用いて，d 後の視聴数，もしくは現在から d 後までの累積視聴数が高人気の定義を満たすものを予測する．

表 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	高人気=H 低人気=L
	スロット 1	スロット 2	$\cdots$	スロット Y		
abcdefghijk	1.0	0.5	$\cdots$	0.4	5	H
lmnopqrstuv	0.5	0.2	$\cdots$	0.0	3	L
wxyz1234567	0.2	0.8	$\cdots$	0.6	4	H

3.3 単純ベイズ分類器の学習と予測の方法

本報告では，将来にわたって高人気を維持する動画の定義として以下の二つの場合を考え，それぞれの場合について単純ベイズ分類器を適用する．

- 定義 1：予測に使用する視聴数データの期間から d 日後の 1 日の視聴数が学習データに含まれる全動画の上位 1%の動画
- 定義 2：予測に使用する視聴数データの期間の翌日から d 日後までの d 日間の累積視聴数が学習データに含まれる全動画の上位 1%の動画

この定義にあてはまるものを「将来にわたって高人気を維持する動画」，あてはまらないものを「高人気を維持しない動画」として二つのカテゴリを用意する．3.1 の入力 n 次元の変数 $F_1, \cdots, F_n$ には，2.3 節と同様に，予測に使用する視聴数データの期間を Y 時間とすると，各動画（総数を v とする）について最初の Y 時間の視聴数の最大値で各時間の視聴数を割った値の小数点第二位を四捨五入した，0 から 1 の範囲の値を要素に持つ Y 次元の変数に，最初の Y 時間の視聴数の最大値の桁数を加えた $Y+1$ 次元の変数を用意する．

3.4 評価結果

本報告では取得した動画の期間の都合上，データセットの動画 87,830 個のうちランダムに選択した半数を教師データとして学習に用い，$p(C=c)$，$p(F_i=f_i|C=c)$ を算出する．残りの半数を試験データに用い，求めた $p(C=c)$，$p(F_i=f_i|C=c)$ を用いて，3.1 節の式（1）の分類器の関数 classify に適用し，予測を行う．アップロード初期の視聴数の推移パターンと視聴数の絶対値から，長期間にわたり高人気を維持する動画を予測する方法として，教師あり機械学習法の一つである単純ベイズ分類器（NBC: Naive Bayes Classifier）を適用する場合に加えて，視聴数上位選択（VCS: View Count based Selection）の場合を評価して比較する．

- 視聴数上位選択法：単純ベイズ分類器で抽出した動画数と同数の動画を，予測に使用する視聴数データの期間 Y の累積視聴数の多い順に選択

評価には正解率を用いる．正解率は，将来にわたって高人気を維持する動画と予測した動画に対して，実際に将来にわたって高人気を維持する動画の定義を満たす動画の割合である．

$d = 7$，予測に使用する視聴数データの期間を $Y = 3$ 時間としたときの結果を表 2 に示す．予測に使用する視聴数データの期間が 3 時間であるので，将来にわたって高人気を維持する動画の定義は

- 8 日目の日視聴数が学習データに含まれる全動画の上位 1%の動画
- 2 日目～8 日目までの 7 日間の累積視聴数が学習データに含まれる全動画の上位 1%の動画

である．単純ベイズ分類器は両方の定義において視聴数上位選択法を上回っている．初期 3 時間の視聴数データを用いて，7 日後も高い視聴数を維持している動画を予測したいとき，累積視聴数の多い順に選択するよりも，単純ベイズ分類器を用いて推移パターンを用いた予測の方が精度が高いことがわかる．

次に，$d = 14$，予測に使用する視聴数データの期間を $Y = 3$ 時間としたときの結果を表 3 に示す．予測に使用する視聴数データの期間が 3 時間であるので，将来にわたって高人気を維持する動画の定義は

- 15 日目の日視聴数が学習データに含まれる全動画の上位 1%の動画
- 2 日目～15 日目までの 15 日間の累積視聴数が学習データに含まれる全動画の上位 1%の動画

である．単純ベイズ分類器は両方の定義において視聴数上位選択法を上回っている．初期 3 時間の視聴数データを用いて，14 日後も高い視聴数を維持している動画を予測したいとき，累積視聴数の多い順に選択するよりも，単純ベイズ分類器を用いて推移パターンを用いた予測の方が精度が高いことがわかる．

次に，$Y = 3$ 時間のとき，d の値を変えた場合の正解率の変化を，定義 1 の場合を図 17(a)，定義 2 の場合を図 17(b) に示す．定義 1 の場合，単純ベイズ分類器の場合は予測日が先になるにつれ正解率が減少する傾向がある．定義 2 の場合は正解率を高く維持している．

次に，$d = 7$ のとき，Y の値を変えた場合の正解率の変化を定義 1 の場合を図 18(a)，定義 2 の場合を図 18(b) に示す．単純ベイズ分類器で入力の数を増加させると，正解の動画の変化パターンが多くなり，正解率が減少するため，$Y/3$ 時間間隔の視聴数で正規化視聴数を

表 2: アップロード後 3 時間の視聴数データを用いた 7 日後の人気度予測の結果

	正解率	
定義	8 日目の視聴数上位 1%	2〜8 日目の累積視聴数上位 1%
単純ベイズ分類器	0.785	0.956
視聴数上位選択法	0.697	0.860

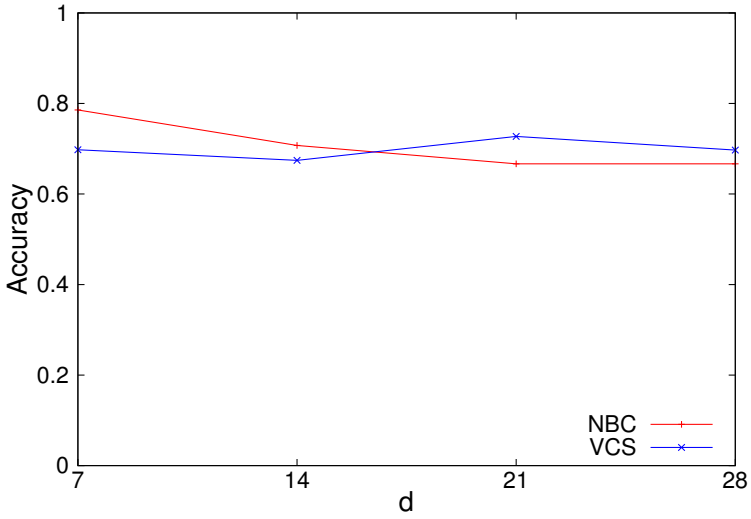
表 3: アップロード後 3 時間の視聴数データを用いた 14 日後の人気度予測の結果

	正解率	
定義	15 日目の視聴数上位 1%	2〜15 日目の累積視聴数上位 1%
単純ベイズ分類器	0.707	0.933
視聴数上位選択法	0.674	0.837

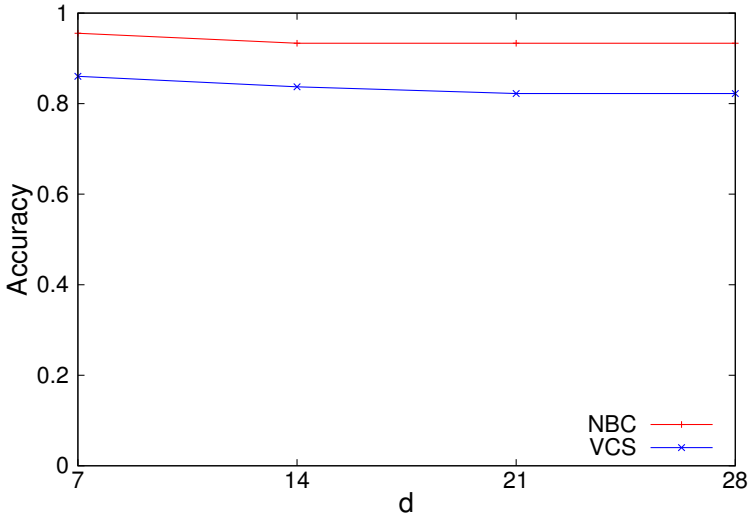
用意し，入力の数を 3 個としている．予測に使用する期間 Y を大きくすると，視聴数の絶対値で見た場合の正解率の方が高いことがわかる．

次に，$d = 14$ のとき，Y の値を変えた場合の正解率の変化を定義 1 の場合を図 19(a)，定義 2 の場合を図 19(b) に示す．$d = 7$ のときと同様に，予測に使用する期間 Y を大きくすると，視聴数の絶対値で見た場合の正解率の方が高いことがわかる．

以上のことから，アップロード後 3 時間の視聴数データを用いて，将来の人気度を予測するとき，視聴数の絶対値で動画を選択するより単純ベイズ分類器の方がより高精度で高人気の動画を予測できることがわかった．これは，アップロード初期は，多くの動画の視聴数が多く，将来的に人気でない動画でも高い視聴数を得ることがあり，単純ベイズ分類器で視聴数の変化パターンを踏まえた予測を行う方が精度が高くなるためと予想される．

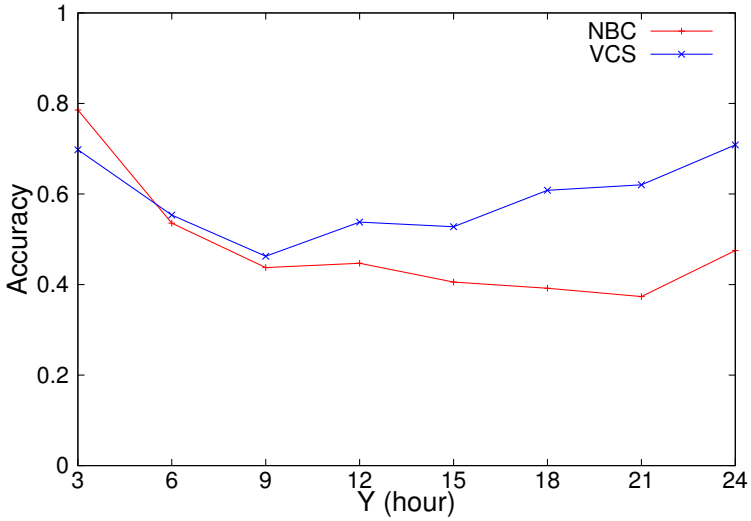


(a) d 日後の日視聴数の予測

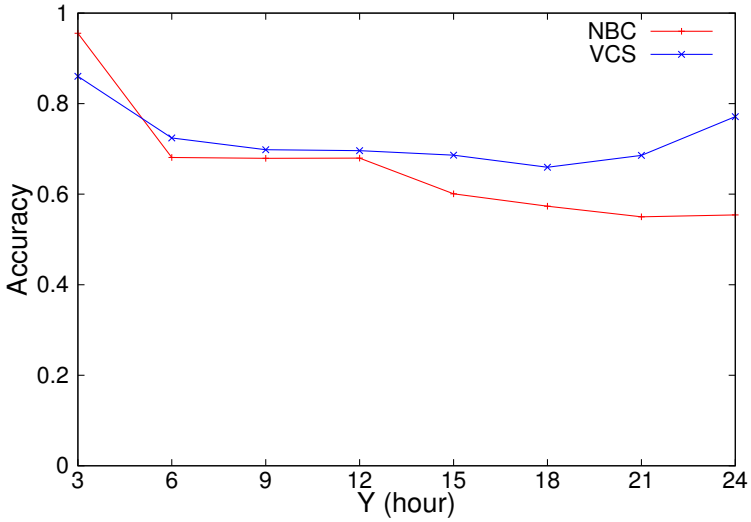


(b) d 日後までの累積視聴数の予測

図 17: $Y = 3$ のときの予測日 d を変えた場合の正解率の推移

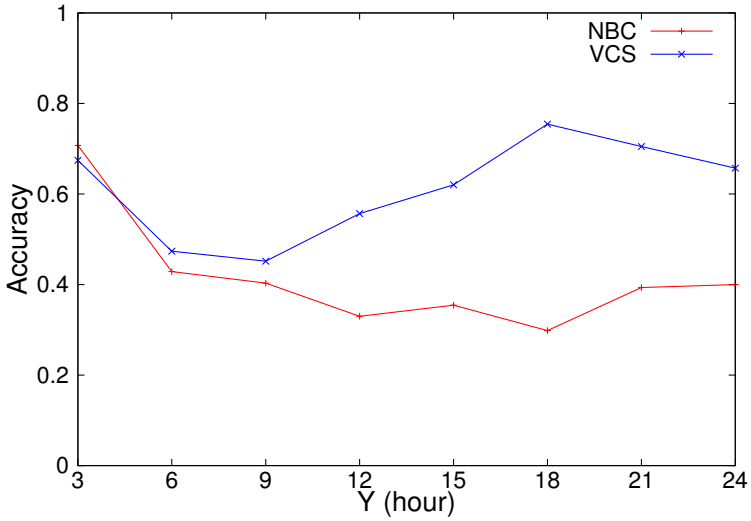


(a) d 日後の日視聴数の予測

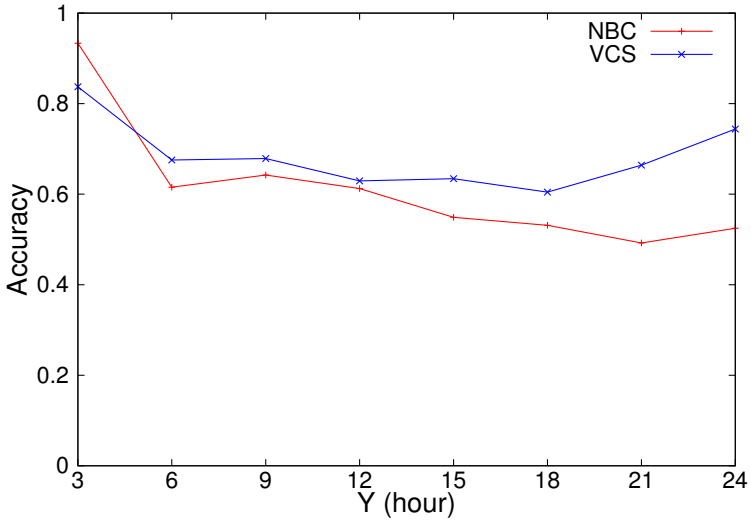


(b) d 日後までの累積視聴数の予測

図 18: $d = 7$ のときの Y を変えた場合の正解率の推移



(a) d 日後の日視聴数の予測



(b) d 日後までの累積視聴数の予測

図 19: $d = 14$ のときの Y を変えた場合の正解率の推移

4　おわりに

本報告では，YouTube 動画の視聴数の時系列データを収集し，視聴数の推移パターンを分析した．その結果，極端に視聴数の大きい少数の動画が存在することが明らかになった．さらに，1 時間単位の視聴数の時系列データを k-means 法でクラスタリングすることで，アップロード初期の視聴数の変化パターンを分析した．その結果，アップロード初期は視聴数が多いがその後視聴数が低下する動画が存在することと，長期間にわたって高い視聴数が維持される動画が存在することがわかった．

また，アップロード初期の視聴数の推移パターンと視聴数の絶対値から，長期間にわたり高人気を維持する動画を予測する方法として，教師あり機械学習法の一つである単純ベイズ分類器を適用した場合の予測精度の評価を行い，アップロード後 3 時間の視聴数データから将来の人気度を予測する場合，初期の視聴数の絶対値のみで予測する場合より，高精度で将来にわたって高人気を維持する動画を予測できることを明らかにした．

今後の課題として，正規化視聴数以外の視聴数データを用いた予測や，アップロードされた大陸や国等のような細かいエリア内における変化パターンの差異の分析・高人気動画の予測や，サポートベクタマシンなどの他の教師あり学習手法による高人気動画の予測，動画視聴数以外のメトリクス（Good 評価数，コメント数，動画ジャンルなど）を用いた高人気動画予測，キャッシュ制御や事前配信，広告ターゲティングに応用した場合の効果の分析などが挙げられる．

謝辞

本報告を終えるにあたり，御多忙の中貴重な御指導，御教授を頂きました大阪大学大学院情報科学研究科の村田正幸教授に心より感謝申し上げます．ならびに本報告の作成にあたりまして，様々な面で終始優しく丁寧に御助言御指導を頂きましたNTTネットワーク基盤技術研究所の上山憲昭先生に心より感謝申し上げます．また，平素から適切な御助言や御指導を頂きました大阪大学大学院情報科学研究科の荒川伸一准教授，大下裕一助教，大阪大学大学院経済学研究科の小南大智助教，ならびに大阪市立大学大学院工学研究科の阿多信吾教授に深く御礼申し上げます．最後に，日頃から様々な質問に答えて頂き，多くの助言，知識を頂きました大岡睦氏，川崎賢弥氏，北川拓氏をはじめとする村田研究室の皆様に深く感謝申し上げます．

参考文献

[1] "YouTube." https://www.youtube.com/.

[2] N. Kamiyama, R. Kawahara, T. Mori, and H. Hasegawa, "Multicast Pre-distribution VoD System," *IEICE transactions on communications*, vol. E96-B, pp. 1459–1471, June 2013.

[3] K. Lerman and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," in *Proceedings of the nineteenth international conference on World Wide Web*, pp. 621–630, Apr. 2010.

[4] "Digg." http://digg.com/.

[5] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: characterizing popularity growth of YouTube videos," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 745–754, Feb. 2011.

[6] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Performance Evaluation*, vol. 68, pp. 1037–1055, Nov. 2011.

[7] G. Gürsun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *Proceedings of IEEE INFOCOM*, pp. 16–20, Apr. 2011.

[8] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, pp. 80–88, Aug. 2010.

[9] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 365–374, Feb. 2013.

[10] G. Chatzopoulou, C. Sheng, and M. Faloutsos, "A first step towards understanding popularity in YouTube," in *Proceedings of INFOCOM IEEE Conference on Computer Communications Workshops*, pp. 1–6, Mar. 2010.

[11] J. M. Tirado, D. Higuero, F. Isaila, and J. Carretero, "Multi-model prediction for enhancing content locality in elastic server infrastructures," in *Proceedings of the*

eighteenth International Conference on High Performance Computing, pp. 1–9, Dec. 2011.

[12] Y. Kitade, "Analyzing popularity dynamics of YouTube content and its application to content cache design," Master's thesis, Graduate School of Information Science and Technology, Osaka University, Feb. 2015.

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

[14] "YouTube Help." https://support.google.com/youtube/answer/2991785/.