社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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〒 565-0871 大阪府吹田市山田丘 1-5
†† 大阪市立大学 大学院工学研究科〒 558-8585 大阪府大阪市住吉区杉本 3-3-138
E-mail: †{t-tanaka,murata}@ist.osaka-u.ac.jp, ††ata@info.eng.osaka-cu.ac.jp

あらまし　近年，YouTube に代表される UGC (User Generated Content) の共有および視聴が非常に活発である．UGC の適切な制御は，サービスにおけるメディア広告の配置等のマーケティング用途だけでなく，コンテンツキャッシュなどの制御によるネットワーク資源の効率利用にも大きく影響を与える．特に，コンテンツ視聴パターンのより正確な把握と予測は，より積極的な最適化制御を実現するためには非常に有効とされており，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推移パターンを予測し，将来も高い人気が継続する動画を予測することが有効である．本稿では UGC 共有サービスとして代表的な YouTube において，各動画の視聴数を実測し，k-means 法を用いたクラスター分析することで，各動画の人気推移パターンを分類する手法を提案し，各推移パターンの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，将来において高い人気が継続すると予想される動画を単純ベイズ分類器により判別する手法を提案する．実測されたデータにもとづく検証の結果，アップロードから 3 時間経過した視聴数推移データを用いた場合，アップロード初期の推移パターン傾向を考慮することで，人気判別精度が 10% 向上することを明らかにした．
キーワード　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

Analysis and Prediction of Popularity Dynamics of User Generated Contents

Tatsuya TANAKA†, Shingo ATA††, and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka, 565-0871, Japan
†† Graduate School of Engineering, Osaka City University
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka-shi, Osaka 558-8585, Japan
E-mail: †{t-tanaka,murata}@ist.osaka-u.ac.jp, ††ata@info.eng.osaka-cu.ac.jp

Abstract　In recent years, social networking services such as YouTube, which share UGC (User Generated Contents) have become much attracted. An efficient control of UGC is one of important role to achieve an optimized placement of advertisements for end users, and/or content-aware caching control for improve the utilization of network resources. To this end, it is effective to forecast the future popularity of the content as early as possible, so that we can take a proactive action to highly popular contents. In this paper, we propose a method to classify time dependent variations of popularity (popularity patterns) of UGCs by using k-means clustering, and analyze tendencies led by popularity patterns. We then propose a method to identify UGCs which are expected to be popular in future, by taking both the initial part of popularity patterns and actual counts of downloads into consideration. Our experimental results show that the accuracy of identification of popular UGCs can be increased around 10% by considering the initial part of popularity patterns.
Key words　YouTube, popularity prediction, view count transition pattern, K-means, Naive Bayes Classifier

1. はじめに

近年，ユーザ生成コンテンツ (UGC; User Generated Content) への注目が高まっており，YouTube [1] や Instagram など，ソーシャルネットワークを通じた共有サービスの利用も非常に活発である．UGC の共有および再生をはじめとする動画

コンテンツの配信は，特に遅延時間に対する要件が厳しく，円滑な再生のためには十分な帯域の確保や安定した遅延特性の維持などが求められる．再生レイテンシの軽減や，ネットワーク資源の有効利用の観点からサービス事業者はネットワーク上に複数分散配置され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，ユーザに安定した視聴品質を維持するためには，キャッシュ性能に対する効果が高いと考えられるコンテンツを優先してキャッシュするなどの制御が必要である．コンテンツキャッシュにおける重要な制御の一つにキャッシュの置きかえアルゴリズムがある．これまで置きかえアルゴリズムとして LRU (Least Recently Used) がたびたび用いられてきたが，UGC の登場によりコンテンツの種類数が飛躍的に増大した結果，個々のコンテンツに対するアクセスパターンが多様化してきている．このため，LRU などコンテンツに依存しない画一的な制御手法では，必ずしも最適なキャッシュ制御が実現できない可能性がある．このため，コンテンツごとのアクセスパターンを把握し，将来の人気度を予測したコンテンツキャッシュ制御手法が有効であると考えられる．

特に，動画配信サーバやネットワークのピーク時負荷を抑制するためには，平常時において将来頻繁なアクセスが予測される動画について，事前に配信することが負荷分散の観点から効果的であると考えられる．そのためには人気度が上昇すると思われるコンテンツの同定を高い精度で行う必要がある．[2].

一方，多くのソーシャルネットサービスにおいて，UGC の投稿者は広告の掲示により，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，サービス事業者側においても，高い視聴数が見込めるコンテンツを効果的に配置し，「注目リスト」などを活用したコンテンツアクセス最適化を行うことで，より効果的な広告マーケティングを行うことができる．その結果，サービス事業者への広告料収入の増加も期待できる．

以上のように，高い視聴数が期待される人気コンテンツを早期に判別し，人気コンテンツを効果的に配信することは，ネットワーク資源の有効利用，利用者の QoE (Quality of Experience) 向上，マーケティングなどさまざまな場面において有効であると考えられるため，人気コンテンツを早期かつ高い精度で識別できる手法の確立が求められる．

これらの背景を踏まえ，本稿では UGC のアクセスパターンにもとづく人気コンテンツの判別を目標とし，UGC の人気度推移の分析，およびそれを利用した人気度予測の手法をそれぞれ提案する．まず，代表的な UGC 共有ソーシャルネットサービスである 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 [3] では時間粒度を 1 日とした場合の視聴数の推移パターンを分析し，高い視聴数が長期間継続する動画の集合を抽出する手法が提案されているが，ネットワーク制御の観点からはより時間粒度が細かい予測が求められる．そこで本稿では，1 時間単位という微細化した時間粒度による時系列データを収集し，それらの推移パターンを分析する．ここでは，k-means 法による推移パターンの分類を行い，細かい時間粒度において視聴数がどのように推移するか，推移パターンごとの傾向を明らかにする．この結果にもとづき，初期の 1 時間ごとの視聴数の推移パターンと視聴数の絶対値から，将来において高い人気が継続すると思われる動画を単純ベイズ分類によって識別する手法を提案する．実測データによる評価の結果，初期の推移パターンを考慮した識別手法では，人気度の高いコンテンツの識別精度を約 10%向上できることを示す．以降，本稿の構成は以下の通りである．まず，2. において UGC の視聴数分析に関連した研究について述べる．次に，3. では，YouTube における 1 時間ごとの視聴数推移の計測および分析手法について述べ，クラスタリングによる推移パターンごとの傾向を明らかにする．そして，4. では，3. の結果にもとづき，高人気のコンテンツを識別する手法を提案し，その効果について述べる．最後に 5. でまとめと今後の課題について述べる．

2. 関 連 研 究

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 [4] はソーシャルニュースサイト Digg [5] 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 [6] 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 [7] 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 [8] 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推移パターンは，長期間一定以上のアクセスが継続するものと単発的の大きく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 [9] 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 [10] は文献 [9] の結果において，アップロードからの経過日数に対する累積視聴数が同じであっても，予測対象日の累積視聴数が大きく異なる動画が存在する点に着目し，アップロードからの人気度の推移パターンから，任意の予測対象日までの累積視聴数を線形回帰モデルで予測することで，予測精度が向上できることを示している．

しかしながら，これらいずれの研究においても，日単位より細かい時間単位の人気度の推移分析および予測についてはあまり検討がなされていない．すでに述べたとおり，人気コンテンツを考慮したネットワーク制御，コンテンツ配置，負荷分散等については，日単位の予測では粒度が荒く適切な制御が行えな

い可能性がある．特に通信トラヒックは 1 日における変動が重要であり，1 時間単位での人気コンテンツの予測が非常に効果的であると考えられる．

3. YouTube 視聴数の測定と分析

3.1 視聴数データ測定方法

本稿では，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，YouTube が提供している API (YouTube Data API version3.0) [11] を用い，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するために独自に開発した収集プログラムを使用して測定を行った．データの取得期間は 2015 年 10 月 14 日から 2015 年 12 月 16 日で，この間において 1 か月以上の期間存在し，かつ視聴数が継続して取得できた動画を対象とする．本章ではこれらの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．

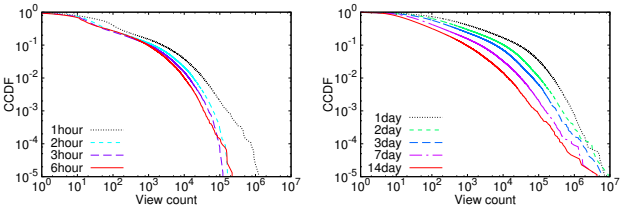
- アップロード 8 日経過時での累積視聴数が 10 以上
- 1 日ごとの視聴数が 5 以下の日が 7 日間継続しない
- 1 時間ごとの視聴数がマイナスにならない

また，YouTube では，動画が不適切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，不適切な方法であると判断された動画の視聴数の修正を行う場合がある [12]．そのため累積視聴数がある時点において減少し，1 時間ごとの差分視聴数がマイナスになる場合があり，これらについても分析対象外とする．

3.2 視聴数分布の時間推移

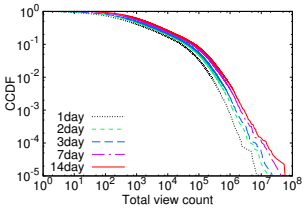
本節では，まずアップロードから一定期間後 (1 時間，2 時間，3 時間，6 時間) の 1 時間の視聴数の累積補分布を図 1(a)，さらに期間を日数単位にしたもの (1 日，2 日，3 日，7 日，14 日) の 1 日の視聴数の累積補分布を図 1(b) に示す．どちらのグラフも両対数グラフで，裾野の部分が線形に近いカーブとなっている．これは極端に視聴数が多い，少数の動画が存在していることを示している．また図 1(a) より，アップロード直後の 1 時間の視聴数が最も大きい．これはアップロード直後は SNS (Social Networking Service) 上などで，「新着動画」のカテゴリなどによって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 1(b) 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 1(c) 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 2(a)，アップロードから 1 日後の視聴数の累積補分布を図 2(b)，アップロードから 7 日後の視聴数の累積補分布を図 2(c) 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，



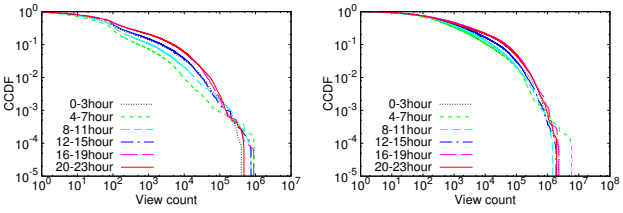
(a) アップロードから 1，2，3，6 時間後の 1 時間の視聴数の CCDF

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



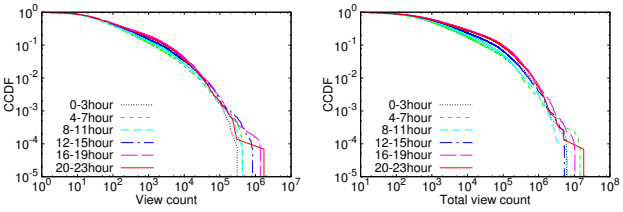
(c) アップロードから 1, 2, 3, 7, 14 日間の累積視聴数の CCDF

図 1　データセットの視聴傾向



(a) アップロードから 1 時間後の視聴数の CCDF

(b) アップロードから 1 日後の視聴数の CCDF



(c) アップロードから 7 日後の 1 日の視聴数の CCDF

(d) アップロードから 7 日間の累積視聴数の CCDF

図 2　アップロード時間帯別の視聴傾向

時間が経過するにつれて，その影響は低下し，むしろ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 2(d) に示す．図 2(c) と同様の傾向が確認できる．

3.3 クラスタリングによる視聴数推移パターン分析

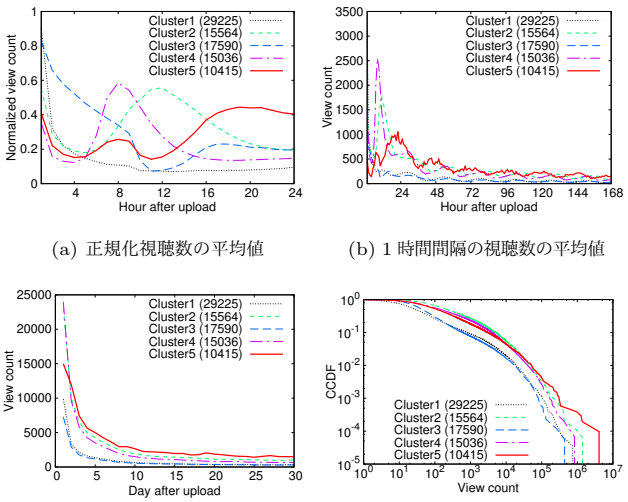
アップロードされてから最初の n 時間における各時間の視聴数に対し，非階層型クラスタリング法の代表的な手法である k-means 法を用いて，YouTube の各動画の人気度の推移パターンに関する傾向を分析する．

文献 [3] では 1 日の粒度の視聴数の推移パターンを 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推移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正規化する．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，推移パターンが似ている動画を分類する．

クラスタ数 5 とし，アップロードから 24 時間までの視聴数推移データを用いてクラスタリングを行った結果を図 3 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間における正規化視聴数の平均値を図 3(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 3(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利用者の視聴が多いため，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，利用者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に視聴数が大きくなるためと考えられる．次に，アップロードから 30 日経過までの日ごとの平均視聴数を図 3(c) に示す．図 3(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 3(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推移パターンをもつ動画がその後，将来におい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 とし，アップロードから 72 時間までの視聴数推移データを用いてクラスタリングを行った結果を図 4 に示す．各クラスタのアップロードからの経過時間における正規化視聴数の平均値を図 4(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 4(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 日経過までの日ごとの平均視聴数を図 4(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が長期間維持され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 4(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推移パターンをもつ動画がその後，将来において



(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 1 日の視聴数の CCDF

図 3　アップロードから 24 時間までの 1 時間ごとの視聴数推移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

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，視聴数が継続す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推移パターンの種類が多くなるため，一つのクラスタにさまざまな推移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推移パターン，初期の正規化視聴数の減少が小さく将来において人気を継続する推移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，4. で，将来において高い人気が継続すると予想される動画を単純ベイズ分類器により判別する場合の予測精度を評価する．

4. 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，将来も視聴数が継続す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．

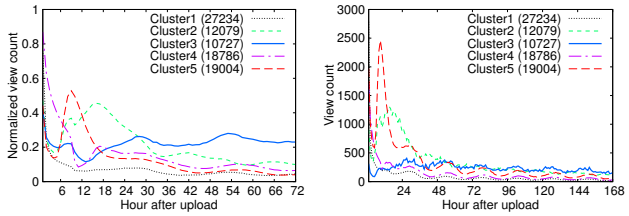
4.1 単純ベイズ分類器の概要

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．学習データから，入力 $F_1, \cdots, F_n$ が与えられたとき，カテゴリ C に分類される確率を計算する．学習データから計算した確率に基づいて，試験データに対して分類カテゴリを決定する．この分類器を関数 classify と以下のように表せる．

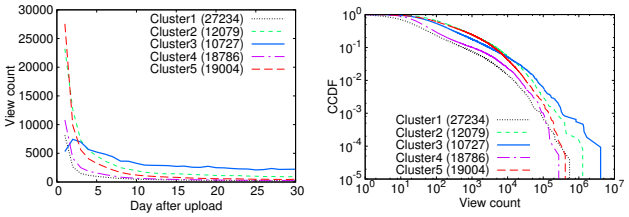
$$\mathrm{classify}(f_1, \cdots, f_n) = \arg\max_c p(C=c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

4.2 高人気動画予測手法

単純ベイズ分類器を用いた高人気動画予測の全体の処理サイクルを図 5 に示す．予測を行う時刻を H，予測に使用するアッ



(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 CCDF

図 4　アップロードから 72 時間までの 1 時間ごとの視聴数推移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

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	高人気=H
	スロット 1	スロット 2	···	スロット Y	視聴数の桁数	低人気=L
abcdefghijk	1.0	0.5	···	0.4	5	H
lmnopqrstuv	0.5	0.2	···	0.0	3	L
wxyz1234567	0.2	0.8	···	0.6	4	H

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．

（1） 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．

（2） 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 時間ごとの視聴数を取得する．

（3） 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．

（4） 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 期間の視聴数推移データを用いて，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．

（5） 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 日後の視聴数，もしくは現在から d 日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．

4.3　単純ベイズ分類器の学習と予測の方法

本稿では，将来において高い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．

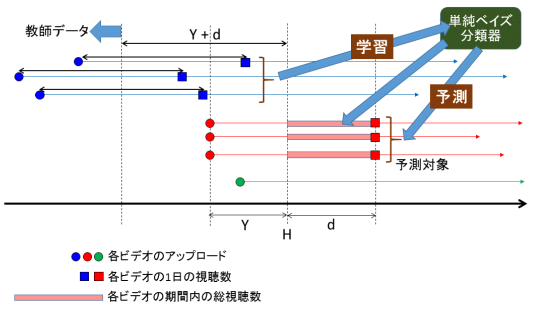


図 5　単純ベイズ分類器を用いた高人気動画予測の処理サイクル

- 定義 1：予測に使用する視聴数推移データの期間から 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推移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「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．4.1 の入力 n 次元の変数 $F_1, \cdots, F_n$ には，3.3 と同様に，予測に使用する視聴数推移データの期間を 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．

4.4　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，残りの半数を試験データに用いて，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，将来において高い人気が維持され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，視聴数上位選択法 (VCS: View Count based Selection) の場合を評価して比較する．視聴数上位選択法は，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推移データの期間 Y の累積視聴数の多い順に選択する方法である．評価には適合率を用いる．適合率は，将来において高い人気が継続する動画と予測した動画に対して，実際に定義を満たす動画の割合である．

予測に使用する視聴数推移データの期間を $Y=3$ 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推移データの期間が 3 時間のため，将来において高い人気を継続する動画の定義は，$d=7$ のとき，

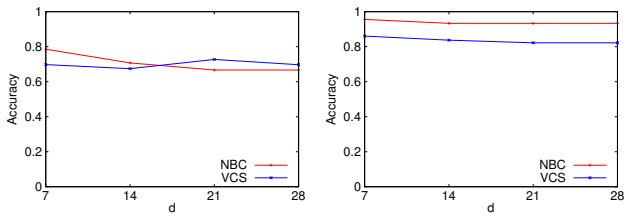
- 8 日目の 1 日の視聴数が学習データに含まれる全動画の上位 1%の動画
- 2 日目～8 日目までの 7 日間の累積視聴数が学習データに含まれる全動画の上位 1%の動画

$d=14$ のとき，

- 15 日目の 1 日の視聴数が学習データに含まれる全動画

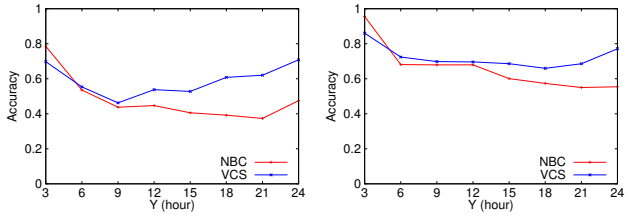
表 2　アップロード後 3 時間の視聴数推移データを用いた人気度予測の結果

	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



(a) 1 日の視聴数の予測　　(b) 累積視聴数の予測

図 6　$Y = 3$ のときの予測日 d を変えた場合の適合率の推移



(a) 1 日の視聴数の予測　　(b) 累積視聴数の予測

図 7　$d = 7$ のときの Y を変えた場合の適合率の推移

の上位 1%の動画

- 2 日目〜15 日目までの 14 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．

$d = 7$，$d = 14$ のとき，初期 3 時間の視聴数推移データを用いて，高い視聴数を維持している動画を予測する場合，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適合率の変化を，定義 1 の場合を図 6(a)，定義 2 の場合を図 6(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適合率が減少する傾向がある．定義 2 の場合は適合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適合率の変化を定義 1 の場合を図 7(a)，定義 2 の場合を図 7(b) に示す．単純ベイズ分類器で入力の数を増加させると，正解の動画の推移パターンが多くなり，適合率が減少するため，$Y/3$ 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適合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 時間の視聴数推移データを用いて，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推移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．

5. ま と め

本稿では，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推移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，将来におい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，将来において高い人気が継続する動画を単純ベイズ分類器で判別する場合の予測精度の評価を行い，アップロードから 3 時間経過した視聴数推移データを用いた場合，初期の視聴数の絶対値のみで予測する場合よりも，高精度で予測できることを明らかにした．今後の課題として，キャッシュ制御や事前配信，広告ターゲティングなどに応用した場合の効果の分析などが挙げられる．

謝　　辞

本研究開発は総務省 SCOPE (受付番号 165007007) の委託を受けたものである．本稿執筆にあたり，適切な御助言を頂いた NTT ネットワーク基盤技術研究所の上山憲昭博士に感謝申し上げる．

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] 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.
[4] 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.
[5] “Digg.” http://digg.com/.
[6] 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.
[7] 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.
[8] G. Gürsun, M. Crovella, and I. Matta, “Describing and forecasting video access patterns,” in *Proceedings of IEEE INFOCOM*, pp. 16–20, Apr. 2011.
[9] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” *Communications of the ACM*, vol. 53, pp. 80–88, Aug. 2010.
[10] 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.
[11] “YouTube Data API.” https://developers.google.com/youtube/v3/.
[12] “YouTube Help.” https://support.google.com/youtube/answer/2991785/.