Osaka University

Analysis of Popularity Pattern of User Generated Contents and its Application to Content-aware Networking

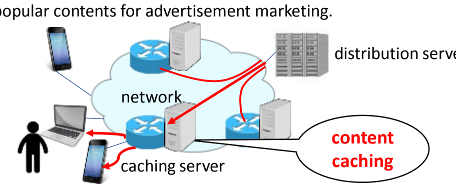
Tatsuya Tanaka[†] Shingo Ata[‡] Masayuki Murata[†]
[†] Graduate School of Information Science and Technology, Osaka University, Japan
[‡] Graduate School of Engineering, Osaka City University, Japan

IEEE GLOBECOM 2016 Workshop on ICNSRA , 8 Dec. 2016 // Washington, DC USA

Osaka University 2

Research Background

- User Generated Contents (UGCs) are becoming popular, which is initiated by social video sharing services such as YouTube.
- **It is effective to forecast the future popular content.**
  - Caching strategy is important in Information Centric Networking.
  - Proactive caching is an effective approach in order to suppress the peak load of the video distribution server.
  - Service provider would like to take a proactive action to highly popular contents for advertisement marketing.



Osaka University 3

Content caching

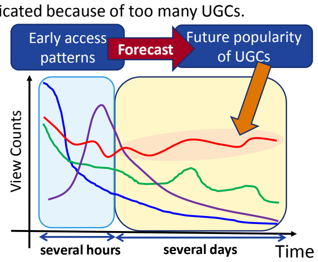
- Content caching is a promising approach to achieve an efficient use of network resources.
  - Many service providers actually utilize a scheme of content cache to improve the end users' Quality of Experience (QoE).
- Cache replacement algorithm is important.
  - Least Recently Used (LRU) is a conventional replacement algorithm.
    - It only focuses on the history of access frequencies.
    - However, it sometimes degrades the overall performance when the distribution is heavily biased.
  - Access frequencies of UGCs heavily depend on their popularity, which may vary significantly in very short term.

Caching strategy should consider future popularity of the content

Osaka University 4

Research Task

- Forecasting the dynamic of UGC popularity is difficult.
  - Popularity pattern is complicated because of too many UGCs.
- There is a correlation between early access patterns and future view counts. [9]
  - It only considered in the unit of 1day.
  - It requires fine grained identification.



[9] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

We aim to forecast future popular contents based on their early access patterns per hour in a short time.

Osaka University 5

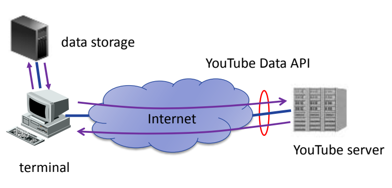
Purposes and procedure of research

- Purposes
  - Analysis of the variation of popularity per hour
  - Proposal of a method to identify future popular contents from the measurement of popularity patterns **around initial phase**
- Procedure
  1. Collect time-series view counts of YouTube videos
  2. Analyze the trend of *popularity patterns* with k-means clustering
  3. Identify a future popular content by using supervised machine learning
     - Apply the Naïve Bayes classifier

Osaka University 6

Collection method of YouTube data

- View counts of recently uploaded YouTube videos
  - We use YouTube Data API version3 [12] to get view counts.
    - (Total 87,830 videos, from Oct.14,2015 to Dec. 5, 2015)
  - Hourly view counts until one week from the initial upload
  - Daily view counts after one week from the initial upload



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

Analyzing popularity pattern by clustering

- We classify popularity patterns with **k-means** clustering.
  - k-means clustering
    - Algorithm of non-hierarchical clustering
- Normalize each hourly view count by the maximum value of hourly view count for first 24 hours
  - Get 24 dimensional vectors having values of $0 \leq s \leq 1$.

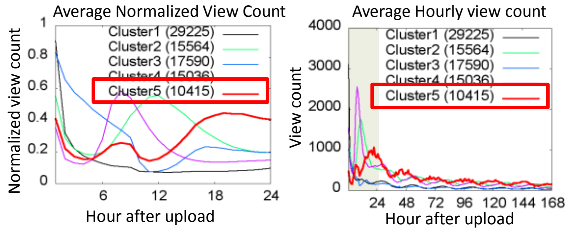
(100, 80, 70, … , 10) Hourly view count → (1.0, 0.8, 0.7, … , 0.1) Normalized view count

Divide by the maximum value

- By using these vectors, classify videos into clusters in which frequency patterns of videos are similar

Results of k-means clustering

- Cluster 5 is the most stable popularity pattern.
  - The normalized view count is not so high just after upload, but it reaches the highest value after 16 hours or later.
  - In other cluster, the normalized view count is decreased sharply.
- Cluster 5 keeps the higher average than others for a long time.



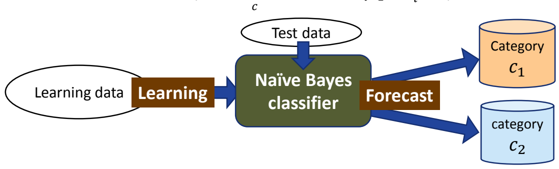
Forecast by supervised learning

- Supervised learning is expected to be effective for popularity prediction.
  - The number of UGC is enormous.
    - Supervised learning can learn various transition patterns.
  - Supervised learning can define popular contents.
- Naïve Bayes Classifier (NBC)
  - A family of simple probabilistic classifiers
  - Despite their naive design and simple assumptions, it has worked quite well in many complex real-world situations.

We apply a Naive Bayes classifier for confirming the efficacy of forecasting by supervised learning.

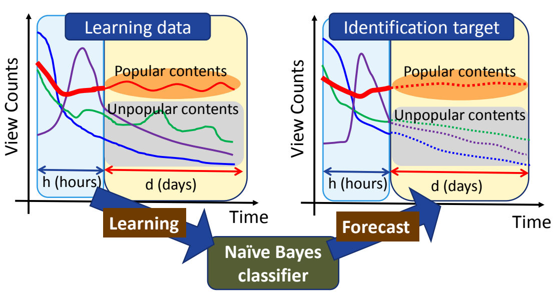
Naïve Bayes Classifier (NBC)

- A kind of supervised learning based on applying Bayes' theorem
  - **Learning**：From the learning data, when input features $F_1, \cdots, F_n$ are given, it calculates a probability of each data to be assigned to a category .
  - **Forecast** : Based on this probability, a classification category is determined for the test data.
  - Function of Naïve Bayes Classifier
    - $\text{classify}(f_1, \cdots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$



Identification procedure

- h (hours)：Period of input data
- d (days)：Target time to identify the popularity



Identification Method of Popular Contents

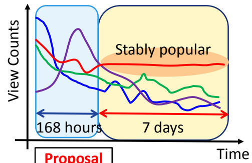
- Output
  - Target1 : **Stably popular contents**
    - The coefficient of variation (CV) of daily view counts in the first d days is lower s% of all videos in the learning data ($s = 1,5,10$).
  - Target2 : **Highly popular contents**
    - Definition1 : Daily view counts in d days are top 1% of all videos.
    - Definition2 : Cumulative view counts in d days are top 1% of all videos.
- Input features
  - Normalized variables obtained by dividing hourly view counts by the maximum hourly view count in first h hours .
  - Digit number of max hourly view count (in the case of Target2)

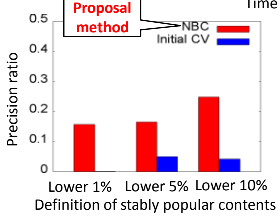
An example of learning data

ID	Normalized view counts				Digit number of max hourly view counts
	Slot 1	Slot 2	…	Slot h	
abcdefghijk	1.0	0.5	…	0.4	5

Evaluation results of stably popular contents

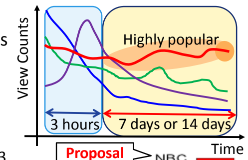
- Identification of stably popular contents by using daily view counts of initial week ($h = 168, d = 7$)
  - Comparison method : selection based on initial Coefficient of Variation (initial CV)
  - Dataset are halved into learning data and test data at random.
- **The precision ratio of the NBC is much higher.**
  - There are many videos which have volatile popularity at initial phase but become stable rapidly.

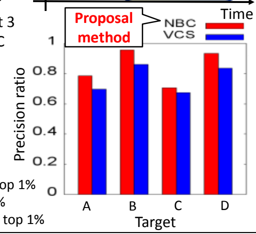




Evaluation results of highly popular contents

- Identification of highly popular contents by using hourly view counts
  - ($h = 3, d = 7,14$)
  - Comparison method : View Count based Selection (VCS)
    - Select the same number of videos of which cumulative view counts in first 3 hours is large in the order as the NBC selected.
- **The precision ratio of the NBC is increased around 10%.**
  - NBC considers popularity pattern.

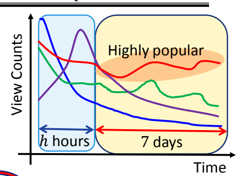




A : Daily view counts in 8 days are top 1%
B : Cumulative view counts in 8 days are top 1%
C : Daily view counts in 15 days are top 1%
D : Cumulative view counts in 15 days are top 1%

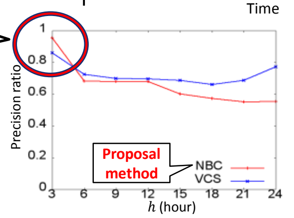
Transition of results by the period of input data

- Transition of precision ratio when we fix $d = 7$ and change the value of h
- Definition of highly popular content
  - Cumulative view counts for 7 days are top 1% of all videos



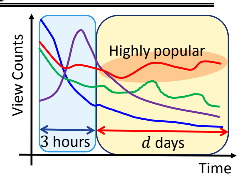
When $h = 3$, the precision ratio of NBC is the maximum.

NBC can identify highly popular contents with high accuracy in a short time.



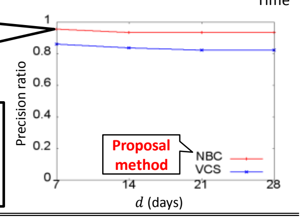
Transition of results by the target time

- Transition of precision ratio when we fix $h = 3$ and change the value of d
- Definition of highly popular content
  - Cumulative view counts for d days are top 1% of all videos



Regardless of the value of d, the precision ratio of the NBC is maintained at high level.

NBC is able to capture the evolution of content popularity, which finally provides high precision in our results.



Summary and future works

- Summary
  - Analysis of the popularity evolution pattern by k-means clustering
    - **There is a popularity pattern that maintains stable view counts.**
    - **Many videos have a popularity pattern which has large view counts just after upload, but decrease sharply.**
  - Application of the supervised learning to identification of popular contents
    - **The identification of Naïve Bayes Classifier which takes the popularity pattern into account grows in performance.**
- Future work
  - This prediction approach will be further evaluated in the control of content caching and advertisement targeting