Prefetching in Video-on-Demand Services based on Recommender Systems

mail

web

dominik@dominikschreiber.com

my name

github

Prefetching in Video-on-Demand Services based on Recommender Systems

amazon.com

code.recommenders

billboard 2014 MUSIC AWARDS

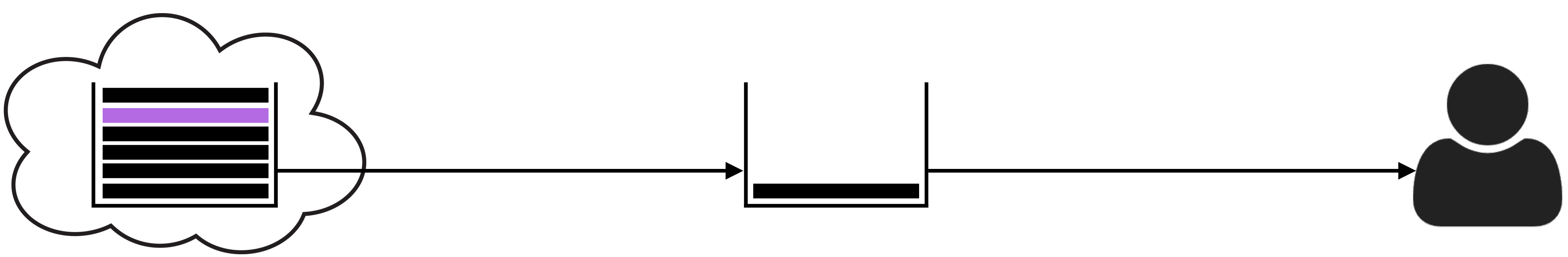
Prefetching in Video-on-Demand Services based on Recommender Systems

NETFLIX

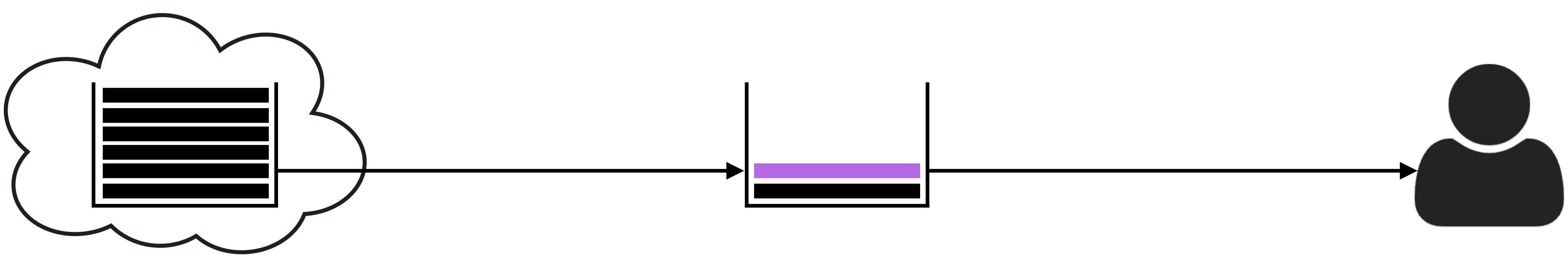
YouTube

OpenLearnWare

Prefetching in Video-on-Demand Services based on Recommender Systems

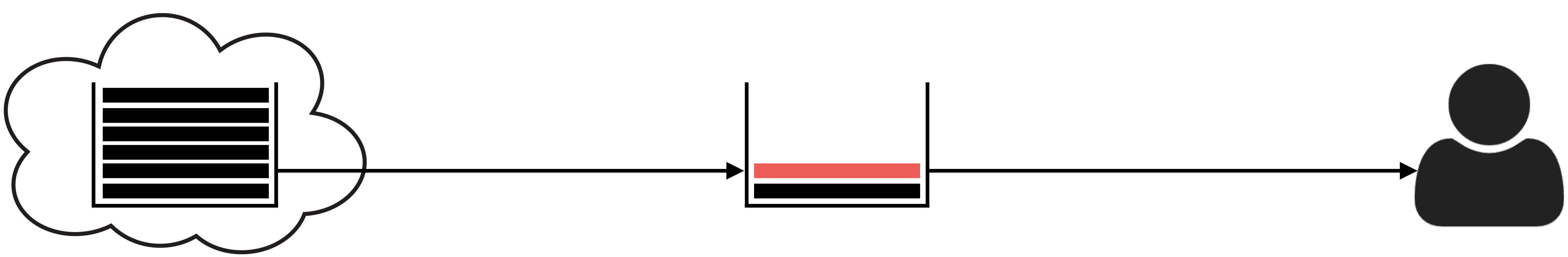


Prefetching in Video-on-Demand Services based on Recommender Systems



Caching

~~Prefetching~~ in Video-on-Demand Services based on Recommender Systems



What to expect?

an overview

the fundamentals

types of VOD services, parameters of interest, probability distributions

recommender systems in VOD services

Netflix’ million-dollar prize, YouTube

research on their impact

in educational context, traditional and user-generated-content services

So, what is VOD?

types of video-on-demand services

NETFLIX

OpenLearnWare

YouTube

General-Purpose

large user base, professional content, high quality, long

Special-Purpose

small audience, single-topic content, focus of first scientific research

User-Generated Content

widespread topics, user participation, lower quality, short

How to study impact?

parameters of interest

File Access Frequency

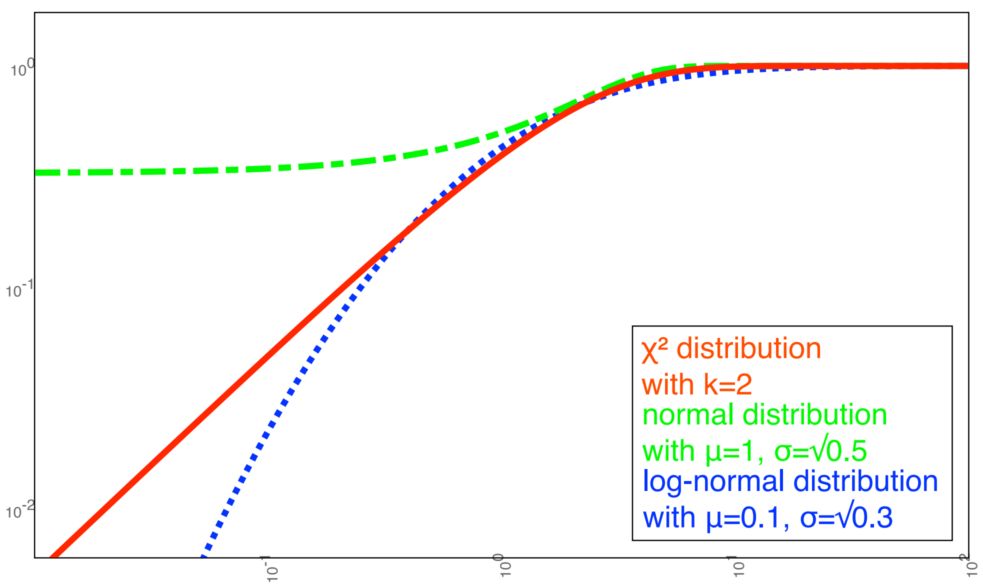
How often has a file been requested in a given period of time?

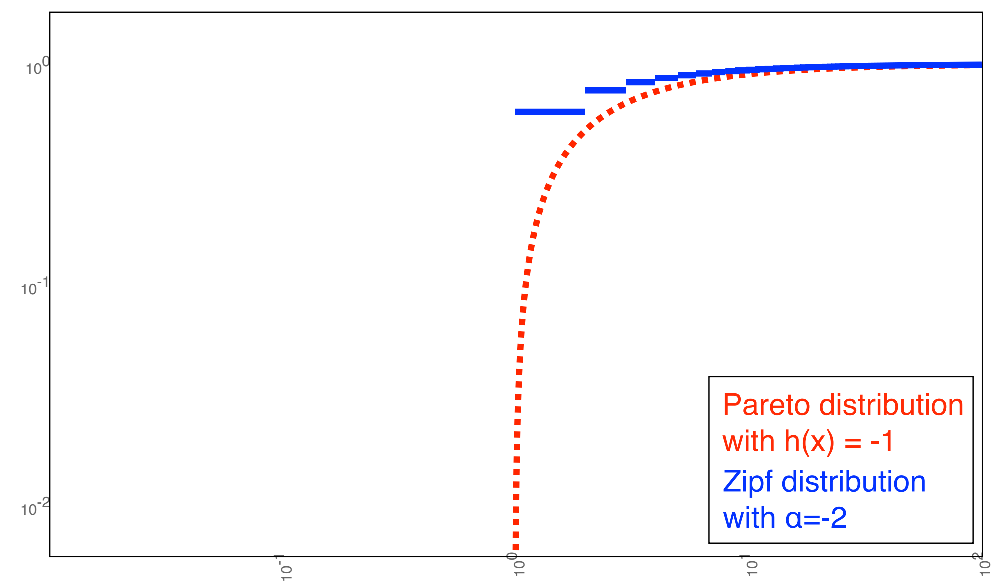
Access Rates

How does file access frequency change over time? Can be normalized to maximum access rates.

What to see there?

probability distributions





Exponentional Distributions

i.e. Gaussian, normal/log-normal, χ^2, Poisson, exponential

$$p_\theta\left(X = x\right) = h\left(x\right)\ exp\left(\theta^\intercal \times T(x) - A(\theta)\right)$$

Power-Law Distributions

i.e. Zipf, Pareto

$$p(X = x) = C\ x^{h(x)}$$

A million dollars for *what*?

Netflix' recommender system

$$
\begin{aligned}
\hat{r}_{ui} = \mu + b_u + b_i + q_i^{\mathsf{T}} &\left(p_u + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_i \right) \\
&+ |R^k(i;u)|^{-\frac{1}{2}} \sum_{j \in R^k(i;u)} (r_{uj} - b_{uj})\, w_{ij} \\
&+ |N^k(i;u)|^{-\frac{1}{2}} \sum_{j \in R^k(i;u)} c_{ij}
\end{aligned}
$$

Neighborhood Model

item neighborhood: find videos similar to the ones already watched

→ *personalized recommendations, less impact on global video popularity*

Latent Factor Model

compare users and videos directly according to inferred factors

→ *search-engine-optimization for latent factors?*

And YouTube?

YouTubes' recommender system

$$C_1(S) = \bigcup_{v_i \in S} R_i$$

$$C_n(S) = \bigcup_{v_i \in C_{n-1}(S)} R_i$$

$$C_{\text{final}}(S) = \left(\bigcup_{i=0}^{N} C_i(S) \right) \setminus S$$

Neighborhood Model

user neighborhood: co-visitation counts for videos

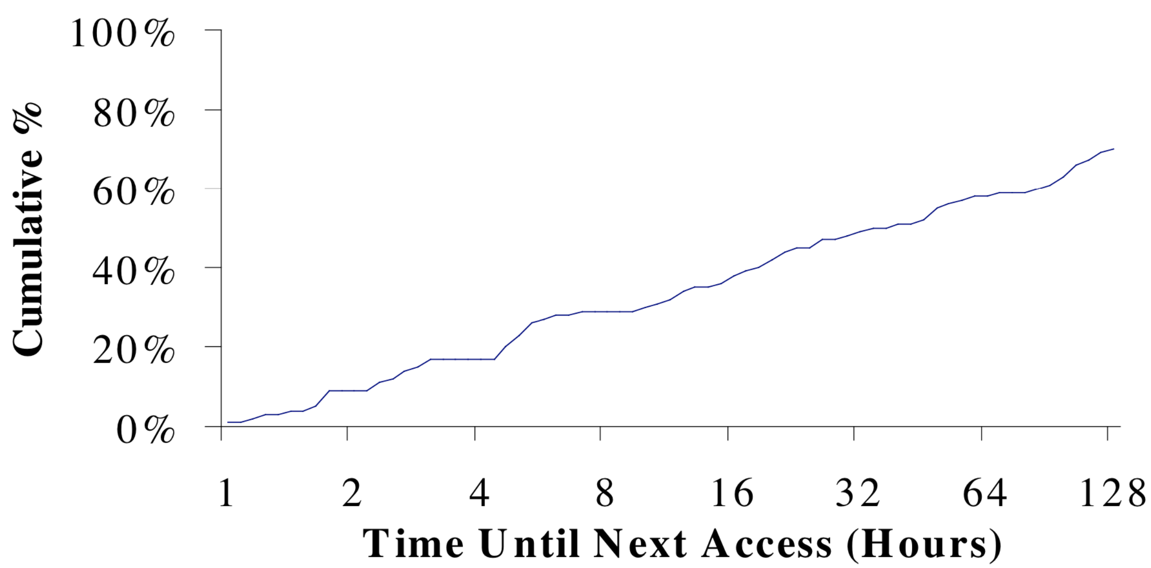
→ *personalized recommendations, less impact on global video popularity*

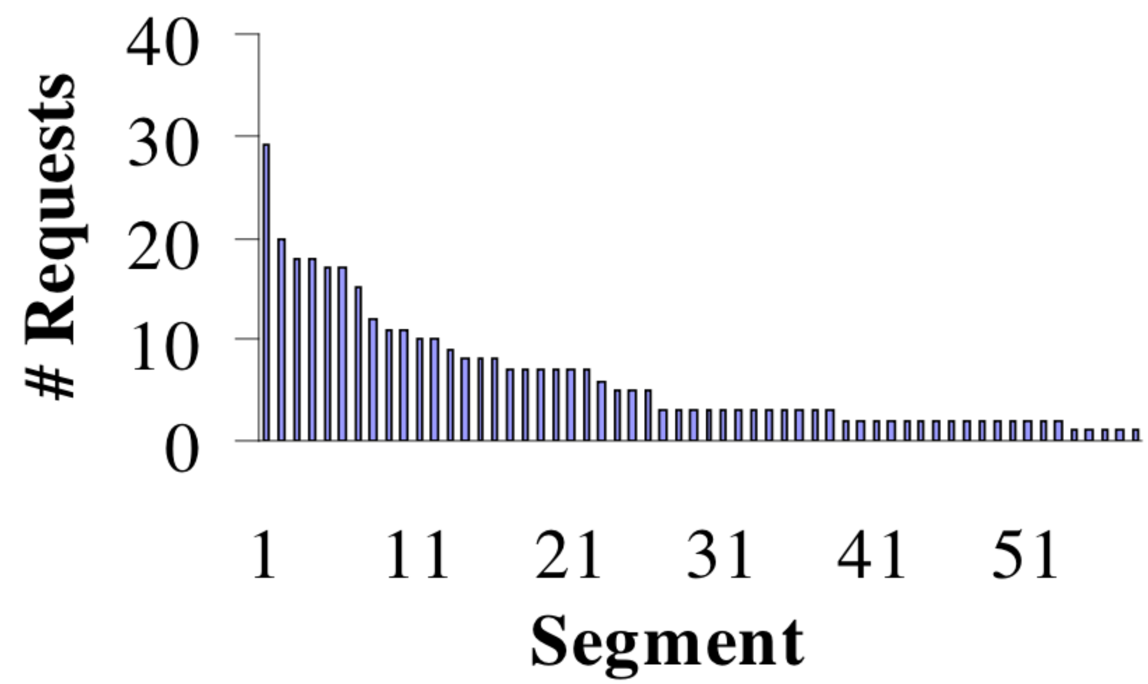
Recommendation Ranking

based on video quality, user specificity and diversification

So, what is the impact?

eTeach & BIBS special-purpose systems studied by Almeida et al.





File Access Frequency

70% of first-time requested videos not requested again in 8h

→ *cache-on-first-hit not useful*

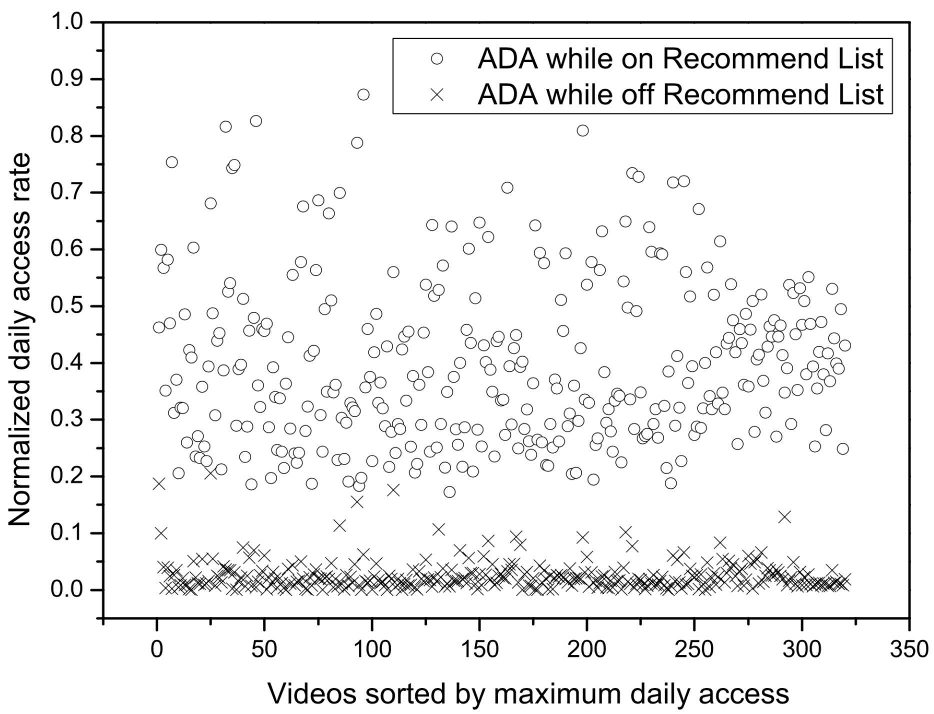
Segment Access Frequency

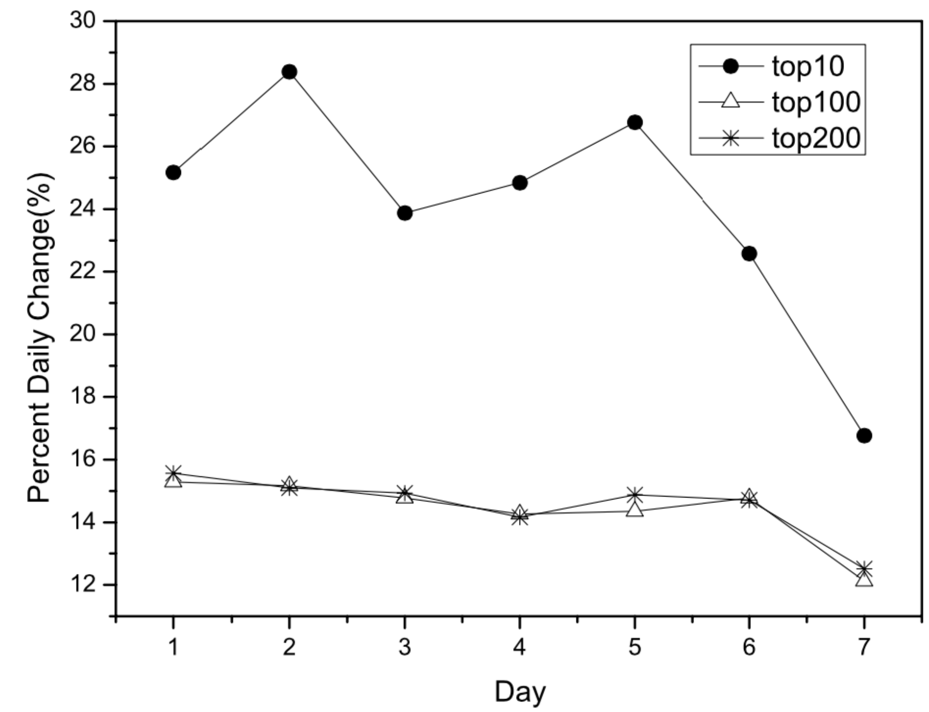
constant for high-ranked videos, first segments more frequently accessed for lower-ranked

→ *prefix caching for lower-ranked videos*

Any direct studies?

PowerInfo general-purpose system studied by Yu et al.





The "Recommended" List

daily access rates of 20-90% while on it but <5% when off it

→ *impact of recommender system clearly visible*

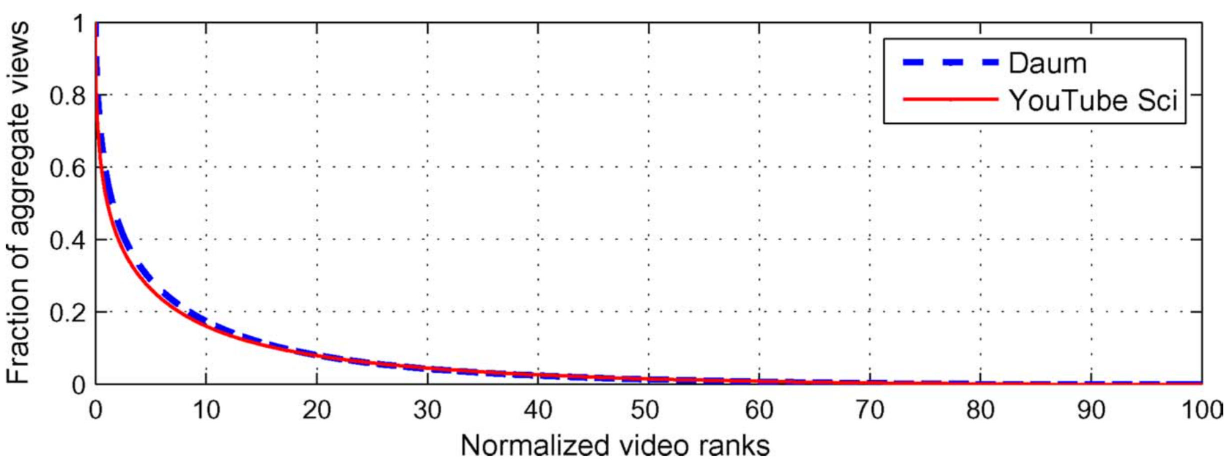
Top 10/100/200 Lists

top 10 constant for single days but totally different for multiple, top 100/200 divergent on single days but stable over multiple days

→ *small but fast "top 10" cache + "top 100" cache for the rest*

But… YouTube?

YouTube & Daum user-generated content systems studied by Cha et al.



Age (x_0)	x_0+5 days old	7 days old	90 days old
2 days old	0.9665 (5185)	0.8793 (3394)	0.8425 (11215)
3 days old	0.9367 (3394)	0.9367 (3394)	0.8525 (9816)

Video Popularity

top 10% of videos account for 80% of views

→ *efficient caching possible*

First Day Determines

popularity on the days 3+ after upload correlates over 90% with popularity before

→ *intelligent caches could cache only videos popular from first day on*

Ok. What to take home?

a conclusion

Recommender Systems

are Collaborative Filtering systems that rely on user/item neighborhood and/or latent factor models

- personalized recommendations weaken the impact on popularity
- global "top *k*" lists are still present and account for large popularity boosts

Caches

can be designed efficiently with the results of the presented research

- cache the *"top 10"* daily
- cache the *"top 100"* in longer term
- cache only *prefixes* of unpopular videos

Fine. Can I have more?

further reading

Recommender Systems In Video-On-Demand Services

Netflix: Yehuda Koren. Factorization Meets the Neighborhood: a Multi-faceted Collaborative Filtering Model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

YouTube: James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293–296. ACM, 2010.

Fine. Can I have more?

further reading

Video Popularity In Video-On-Demand Services

eTeach & BIBS: Jussara M. Almeida, Jeffrey Krueger, Derek L. Eager, and Mary K. Vernon. Analysis of Educational Media Server Work- loads. In *Proceedings of the 11th International Workshop on* Network and Operating Systems Support for Digital Audio and Video, NOSSDAV ’01, pages 21–30. ACM, 2001.

PowerInfo: Hongliang Yu, Dongdong Zheng, Ben Y Zhao, and Weimin Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *ACM SIGOPS Operating Systems Review*, volume 40, pages 333–344. ACM, 2006.

YouTube: Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: Analyzing the world’s largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC ’07, pages 1–14. ACM, 2007.

So Long, and Thanks for All the Fish

Grab A Copy

The paper: dominikschreiber.com/papers/vod.pdf
The slides: dominikschreiber.com/talks/vod.pdf

Let’s discuss it!

*"Strong minds discuss ideas,
average minds discuss events,
weak minds discuss people."*
—Sokrates