IA316

Youtube environment

Thomas Dahmen - Paul Chevalier

I. Environment presentation
II. Experiments & results

Real Youtube

Users = channels

Users can :

- post videos
- watch videos (watch time)
- like / dislike videos
- suscribe to channels
- create playlists and add videos to them

Our Youtube

Users ≠ channels

Channels can post videos

Users can watch videos

Watch time $\in [0, 1]$ (≃ like)

Real Youtube

Videos :

- explicit content (3D tensor)
- implicit content (tags, channel, metadata…)

Our Youtube

Videos = feature vectors, **n dimensional, unit norm**.
Each feature = a content **keyword** (e.g. : humor, rock, sport, …).
Sparse feature vectors = mix of a few keywords.
Same for users : tastes = sparse unit vectors of keywords.

Rewards

Reward = watch time $\in [0, 1]$.
$\rightarrow$ maximize the watch time of recommended videos.

User **u**, video **v**.
Cosine similarity for tuple (**u**, **v**) : s(**u**, **v**).

Reward probability model for (**u**, **v**) :

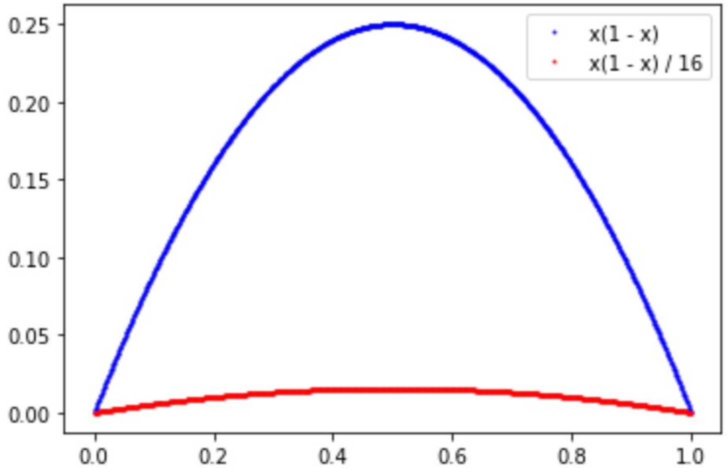
- the higher the similarity, the higher the watch time in average
- extreme similarities (≃ 0 or 1) incur less variance

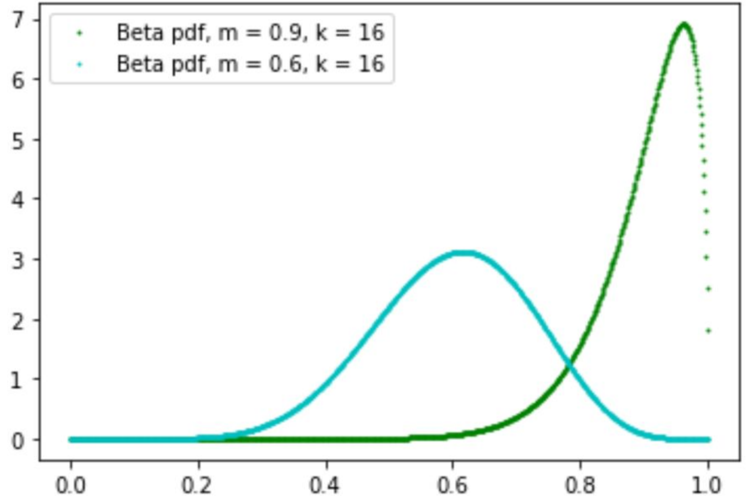
Rewards

Mean = m = s(**u**, **v**)
Var = (m * (1 - m)) / k

→ Reward ~ Beta(a, b)

a = (k - 1) * s(**u**, **v**)
b = (k - 1) * (1 - s(**u**, **v**))





Evolution

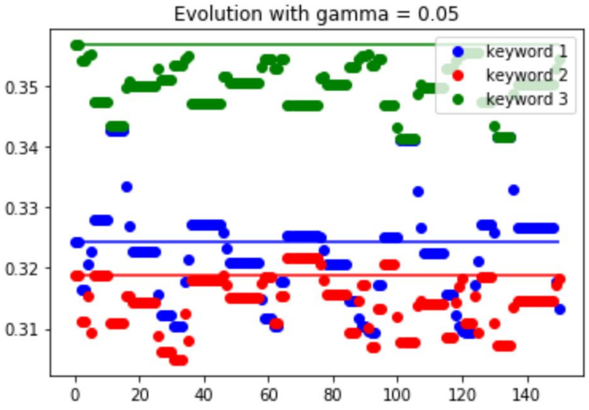
Recommended videos can influence user tastes.
$\rightarrow$ **the environment can be evolutive**

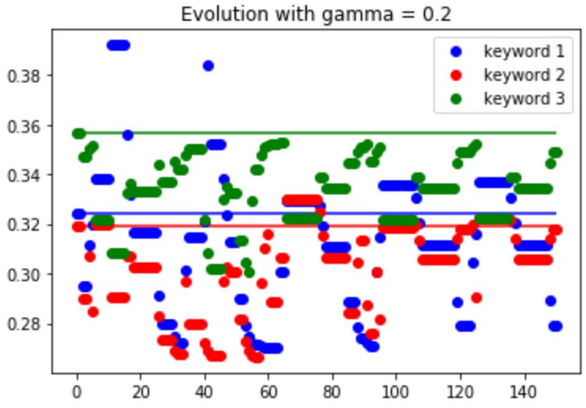
User **u** ($\mathbf{u_0}$ at first), recommended video **v**, watch time **r** $\in [0, 1]$
r ≃ continuous version of "like"
With probability **r**, let **u** change :

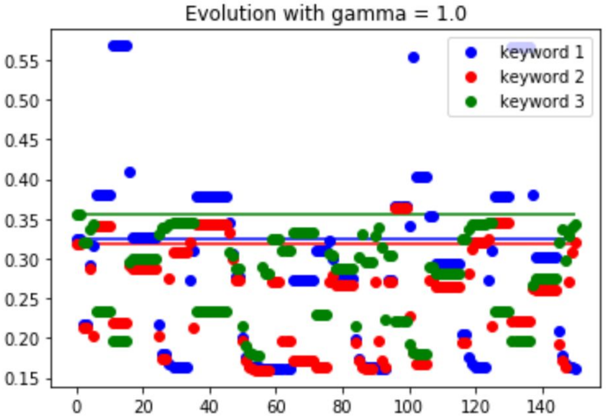
$$\mathbf{u} \leftarrow \mathbf{u_0} + \mathbf{u} + \gamma \times \mathbf{v}$$

γ : influence of the recommended videos

Evolution


Evolution with gamma = 0.05
keyword 1
keyword 2
keyword 3
0.35
0.34
0.33
0.32
0.31
0
20
40
60
80
100
120
140


Evolution with gamma = 0.2
keyword 1
keyword 2
keyword 3
0.38
0.36
0.34
0.32
0.30
0.28
0
20
40
60
80
100
120
140


Evolution with gamma = 1.0
keyword 1
keyword 2
keyword 3
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0
20
40
60
80
100
120
140

Experiments

Agents used:

- Epsilon Greedy (epsilon = 0.1)
- Thompson Sampling
- Q-Learning

In two cases:

- Non evolutive environment
- Evolutive environment

Environment parameters

- 10 users
- 10 channels
- 3 videos per channel
- 100 keywords
- 3 keywords per user
- 3 keywords per video

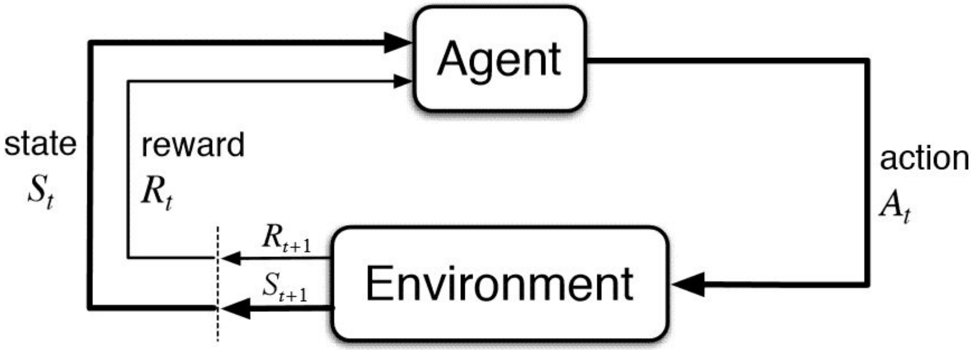
Q Learning

Initialized

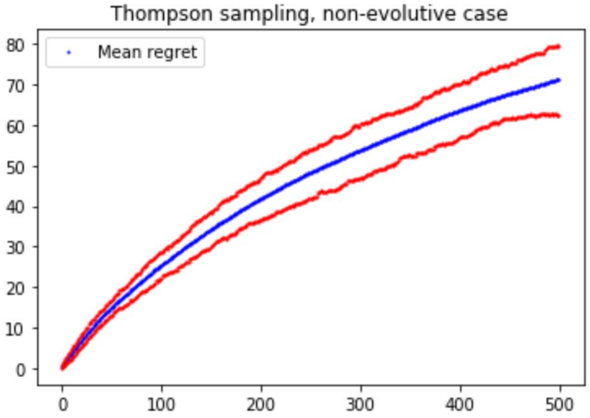
Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	327	0	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	499	0	0	0	0	0	0

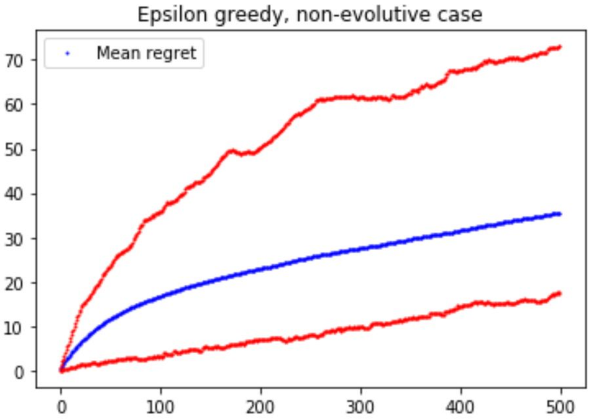
Training

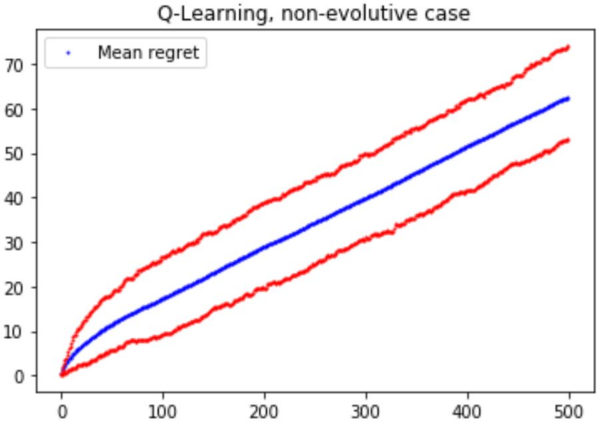
Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	328	-2.30108105	-1.97092096	-2.30357004	-2.20591839	-10.3607344	-8.5583017
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	499	9.96984239	4.02706992	12.96022777	29	3.32877873	3.38230603



Experiments in non-evolutive case

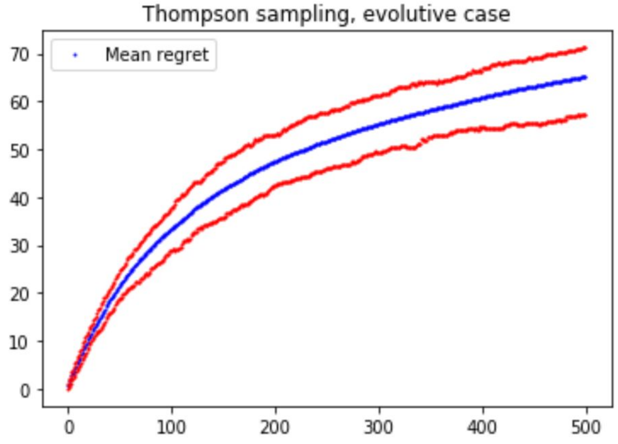


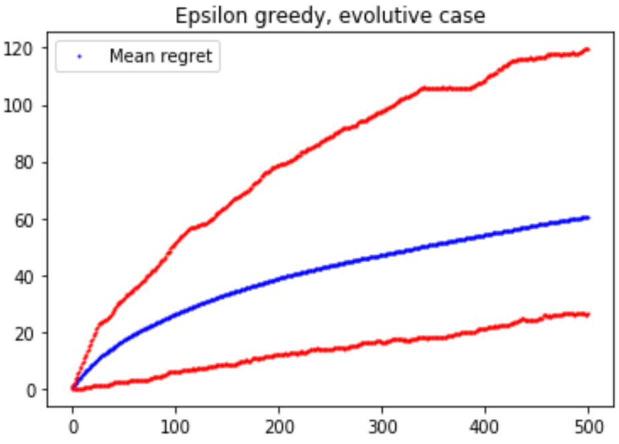


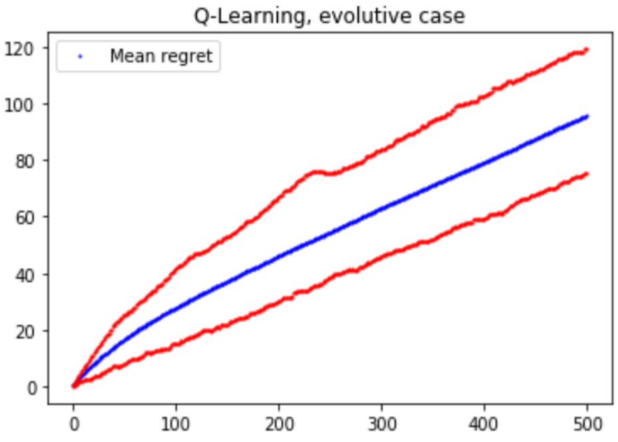


- Time horizon : 500 steps
- Number of simulations : 100

Experiments in evolutive case







- Time horizon : 500 steps
- Number of simulations : 100

Conclusion

- Epsilon Greedy is better in average, but has a high variance
- Thompson Sampling is more stable
- Q Learning is the less satisfying agent
- In average, regret is higher for an evolutive environment

To go further...

- Experiment with a bigger environment (more users and videos)
- Change video/user ratio
- Use collaborative algorithms (such as top videos)
- Use Deep Q Learning