Contenu généré par les utilisateurs : une étude sur DailyMotion†

Yannick Carlinet[1], The Dang Huynh[24], Bruno Kauffmann[1], Fabien Mathieu[24], Ludovic Noirie[24] et Sébastien Tixeuil[34]

[1]*Orange Labs* [2]*Alcatel-Lucent Bell Labs France* [3]*UPMC Sorbonne Universités, IUF* [4]*LINCS*

Actuellement, une large part du trafic Internet vient de sites de *User-Generated Content* (UGC) [1]. Comprendre les caractéristiques de ce trafic est important pour les opérateurs (dimensionnement réseau), les fournisseurs (garantie de la qualité de service) et les équipementiers (conception d'équipements adaptés). Dans ce contexte, nous proposons d'analyser et de modéliser des traces d'usage du site DailyMotion [2].

Une version étendue de cet article a été publiée à TRAC 2012 [3].

Keywords: DailyMotion, Analyse de traces, modélisation

1 Données utilisées

Les données sur DailyMotion ont été collectées à partir de 7 sondes placées au sein du réseau d'accès d'Orange, permettant d'observer 65400 utilisateurs. Pendant quatre mois (février à mai 2010), les requêtes HTTP demandant un fichier flash (`.swf`) au site DailyMotion ont été collectées, soit un total de 4948593 évènements. Chaque évènement est décrit par : l'utilisateur ; le contenu demandé ; la date d'émission de la requête ; le nombre d'octets récupérés. Les données sont anonymisées.

Après une première analyse, les évènements ont été classés en 3 catégories :

- *Lanceur :* avant de récupérer une vidéo, le navigateur récupère un lecteur en flash qui va à son tour lancer une ou plusieurs requêtes (environ 20% des évènements).
- *Vidéo :* requête d'une vidéo autre que la précédente requête du même utilisateur ($\approx 40\%$).
- *Saut :* souvent, un utilisateur demande plusieurs fois de suite le même film. L'interprétation est un saut : l'utilisateur va vers une partie de la vidéo qui n'a pas encore été enregistrée localement ($\approx 40\%$).

Pour affiner l'analyse, les évènements ont été regroupés en sessions utilisateur. Une première idée, se servir des lanceurs pour définir une session, a rapidement été abandonnée : un navigateur peut conserver le lanceur dans son cache, ou à l'inverse un utilisateur peut rafraichir son navigateur, provoquant une nouvelle récupération du lanceur. Nous avons donc préféré ignorer les requêtes de lanceurs et utiliser les périodes d'inactivité pour la classification. Le choix empirique a été de fixer un seuil fixe d'une heure, à partir duquel ont été extraites 567510 sessions.

2 Arrivées des sessions

Pour analyser la distribution des débuts de session, il est naturel de supposer l'absence de corrélation entre les utilisateurs, ce qui suggère une distribution de Poisson [4]. Comme l'intensité moyenne observée est très fluctuante (voir Figure 1), une loi hétérogène est la plus adaptée : on cherche une fonction $\lambda(t)$ (en arrivées par secondes), telle que la probabilité d'avoir k arrivées entre t_1 et $t_2 > t_1$ est donnée par

$$P(k,t_1,t_2) = \mathcal{P}(k,\lambda_{t_1}^{t_2}), \text{ with } \begin{cases} \lambda_{t_1}^{t_2} = \int_{t_1}^{t_2} \lambda(t)dt, \\ \mathcal{P}(k,\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}. \end{cases} \quad (1)$$

†Les travaux présentés ici ont été en partie réalisés au *Laboratory of Information, Network and Communication Sciences* (LINCS, `http://www.lincs.fr`).

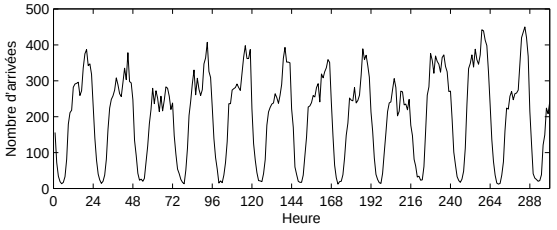


FIGURE 1: Nombre de sessions par heure observées sur quelques jours

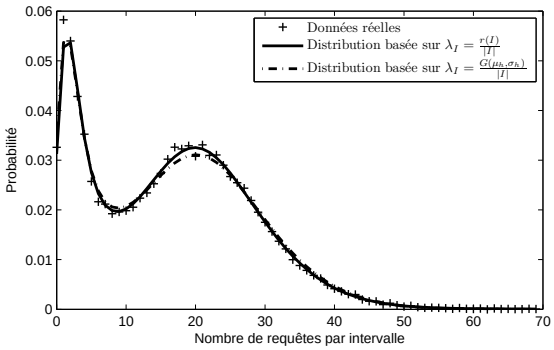


FIGURE 2: Distribution du nombre de débuts de sessions par intervalle de cinq minutes (traces et modèles).

Connaître λ permettrait de générer des arrivées artificielles statistiquement indistinguables des arrivées réelles, mais une estimation parfaite est impossible. Nous proposons donc d'approximer λ par une fonction constante par morceaux : étant donnée une partition en intervalles $\mathbf{I}$ de la période, pour $I \in \mathbf{I}$ et $t \in I$, nous supposons

$$\lambda(t) \approx \lambda_I := \frac{r(I)}{|I|}, \text{ où } |I| \text{ est la durée de } I. \tag{2}$$

Cette approche autorise un découpage hétérogène de la période, mais empiriquement, une partition en intervalle d'une heure a semblé un compromis raisonnable (cf Figure 1).

Les λ_I donnent une bonne idée de la distribution des sessions, mais cela représente beaucoup de paramètres (2880 pour 120 jours de traces), difficiles à extrapoler (par exemple sur une période plus longue). Afin de réduire les paramètres tout en offrant plus de flexibilité, nous proposons d'utiliser un modèle journalier bruité : si $\mathbf{I}_h$ est l'ensemble des intervalles I correspondant à l'heure $[h, h+1]$, pour $I \in \mathbf{I}_h$, nous prenons

$$\lambda_I = \frac{G(\mu_h, \sigma_h)}{|I|}, \tag{3}$$

où $G(\mu_h, \sigma_h)$ est pris selon une loi normale tronquée de paramètres μ_h et σ_h estimés à partir de $\mathbf{I}_h$.

L'équation (3), permet de simuler des arrivées sur une période de temps arbitrairement longue, mais aussi de modifier le comportement journalier à loisir. L'inconvénient majeur de cette simplification est qu'elle efface les corrélations d'une heure à l'autre. Cet effet pourrait être compensé en combinant plusieurs variables aléatoires (par exemple une intensité journalière et une intensité horaire), mais le rapport complexité/gain n'est pas forcément intéressant, comme l'illustre la Figure 2, qui compare la distribution des arrivées pour : la trace réelle ; le modèle par intervalles ; le modèle journalier bruité.

3 Déroulement d'une session

Nous avons d'abord étudié la durée des sessions, extrapolée à partir des dates de la première et de la dernière requête (pour n requêtes, une correction $\frac{n}{n-1}$ est appliquée). Le résultat est représenté Figure 3 :

- 34% des sessions ne comportent qu'une unique requête, et ne peuvent être estimées ;
- la plupart des autres sessions ne durent que quelques minutes ; la durée médiane est de 7 ;
- 25% des sessions dépassent la demi-heure, 13% dépassent l'heure ;
- des sessions longues existent (3.3% de plus de 2 heures, 1% de plus de 3,5 heures).

Pour mieux comprendre le déroulement d'une session, nous avons ensuite utilisé un modèle Markovien : à chaque session S de longueur n correspond une séquence $(e_i)_{1 \le i \le n}$, où $e_i \in \{V, S\}$ est le type d'évènement (vidéo ou saut). Pour chaque séquence $(e_j)_{1 \le j \le k}$ qui est préfixe d'au moins une session, nous avons calculé la probabilité du prochain évènement : vidéo, saut, ou rien (fin de la session). On obtient ainsi une chaîne de Markov qui décrit précisément les séquences observées. L'inconvénient de cette approche exhaustive est le nombre d'états nécessaires, mais nous avons réussi à trouver une réduction dont la distribution des

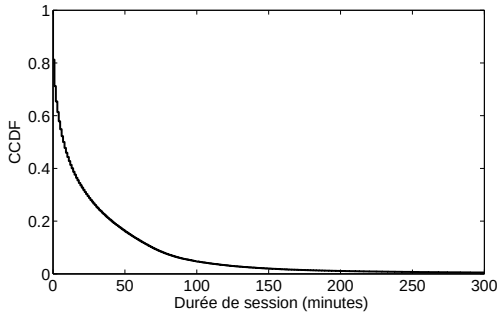


FIGURE 3: Distribution cumulée complémentaire de la durée des sessions

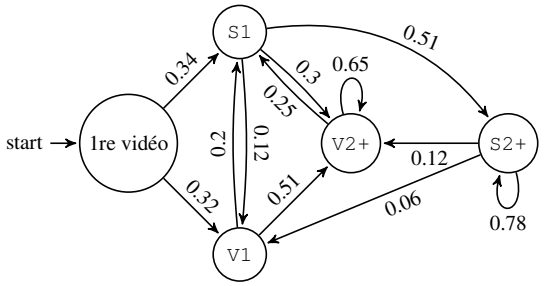


FIGURE 4: Chaine de Markov compacte de session

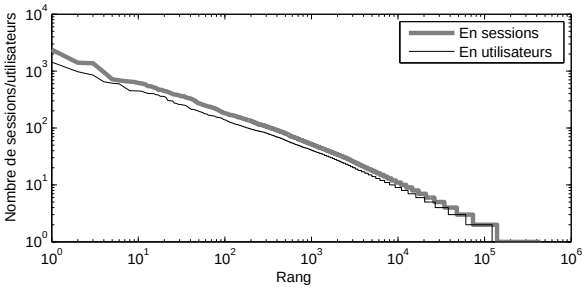
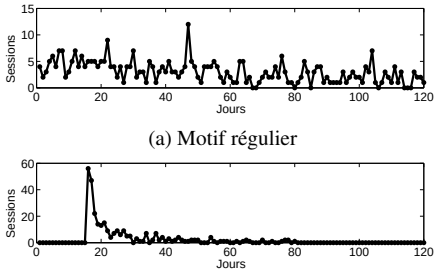


FIGURE 5: Popularité en fonction du rang



(a) Motif régulier

(b) Motif en pic

FIGURE 6: Motifs temporels de popularité

séquences diffère de l'original de moins de 1%. Le résultat, présenté Figure 4, est une chaîne à cinq états (plus la fin de session, implicite).

- Après le lancement de la première vidéo, les trois options (nouvelle vidéo, saut ou fin de session) sont à peu près équiprobables.
- Les états `V1` et `V2+` correspondent à des nouvelles vidéos. `V1` représente souvent (mais pas seulement) le premier changement, et `V2+` les changements ultérieurs. La transition du départ vers `V1` est de 32%, celle de `V1` vers `V2+` de 51% et que la boucle sur `V2+` est de 65% : les chances de regarder une nouvelle vidéo augmentent avec le nombre de vidéos déjà regardées (0, 1, ou plus de 2).
- Les états `S1` and `S2+` correspondent à des sauts. `S1` est (exactement) le premier saut d'une vidéo donnée, et `S2+` les sauts suivants. Comme pour les vidéos, la probabilité de saut augmente en fonction des sauts précédents : entre 20% et 34% pour le premier saut (selon l'état initial), 51% pour le deuxième et 78% ensuite.

4 Popularité des vidéos

La popularité peut se mesurer via divers moyens : en nombre de requêtes brutes (sauts inclus), de sessions ou d'utilisateurs. Au niveau sessions, les données comportent 440000 vidéos distinctes, soit une moyenne de 1,29 session par vidéo. On observe un comportement relativement classique à aile lourde : la vidéo la plus populaire compte 2381 sessions, 338 vidéos comptent plus de 100 sessions, tandis qu'une large majorité (300000) de vidéos n'apparaît qu'une fois. Ces données sont illustrées par la Figure 5.

Un autre aspect important de la popularité est sa répartition temporelle. Nous avons examiné le nombre de sessions par jour des 10000 vidéos les plus populaires (en-deça, le nombre de sessions par vidéo est trop faible), et deux principaux types de profil sont apparus : un motif régulier (Figure 6a), où un nombre relativement constant de requêtes est observé sur un grand nombre de jours ; un motif en pic (Figure 6b), où la grande majorité de sessions est concentrée sur quelques jours.

La répartition des motifs est intéressante : sur les 10000 vidéos, la proportion de motifs en pic est de 15%. Cette proportion monte à 50% pour les 2000 premières vidéos, et à 90% pour les 100 premières. Autrement dit, la majorité des vidéos ont un motif temporel régulier, mais les vidéos les plus populaires ont

un motif en pic. L'observation de cette corrélation nous semble importante, car elle a potentiellement de fortes implications, par exemple pour les performances d'algorithmes de mises en cache du contenu.

5 Travaux connexes

La plupart des travaux connexes aux présents résultats se basent sur le site YouTube, en raison de sa popularité. Parmi les travaux comparables aux nôtres, on peut citer :

- Gill *et al.* [5], qui utilisent la décomposition des requêtes en sessions sur des données de 2008 (le seuil d'inactivité choisi était de 40 minutes).
- Abhari et Soraya [6], qui infèrent le comportement des utilisateurs à partir de données issues du site YouTube (par opposition notre approche est basées sur des mesures passives au niveau utilisateur) ;
- Plissonneau *et al.* [7] utilisent la même approche de mesure que nous, mais sur une période plus courte (35 heures) ;
- enfin, diverses études de popularité ont été faites, souvent mesurées à partir du fournisseur de contenu (voir par exemple [6, 8] pour Youtube).

6 Conclusion

En analysant au niveau utilisateur quatre mois d'usage du site DailyMotion, nous proposons de modéliser les données par la combinaison de 3 comportements élémentaires : des arrivées de sessions suivant un processus de Poisson hétérogène ; un comportement pour chaque session décrit par une chaîne de Markov ; une popularité des vidéos en aile lourde, avec de fortes corrélations temporelles, en particulier sur les vidéos populaires. Si nous avons observé de petites corrélations entre ces trois processus, il semble possible de les négliger en première approximation.

Le résultat est alors une décomposition naturelle permettant de décrire le comportement des utilisateurs, qui ouvre la voie à la création d'un générateur de traces, où les différents éléments (arrivées, sessions, popularité) pourraient de manière indépendante être extraits de traces réelles ou émulés (avec les paramètres initiaux ou en changeant certaines hypothèses).

Références

[1] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere : impact of device and infrastructure synergies on user experience," in *IMC '11*, 2011, pp. 345–360.

[2] Dailymotion, http://www.dailymotion.com.

[3] Y. Carlinet, T. D. Huynh, B. Kauffmann, F. Mathieu, L. Noirie, and S. Tixeuil, "Four Months in DailyMotion : Dissecting User Video Requests," in *TRAC 2012 - 3rd International Workshop on TRaffic Analysis and Classification*, Limassol, Cyprus, Aug. 2012. [Online]. Available : http://hal.inria.fr/hal-00692095

[4] C. Park, H. Shen, J. S. Marron, F. Hernandez-Campos, and D. Veitch, "Capturing the elusive poissonity in web traffic," in *14th IEEE MASCOTS Conference*, 2006, pp. 189–196.

[5] P. Gill, M. F. Arlitt, Z. Li, and A. Mahanti, "Characterizing user sessions on youtube," in *SPIE/ACM Conference on Multimedia Computing and Networking (MMCN)*, Santa Clara, USA, 2008.

[6] A. Abhari and M. Soraya, "Workload generation for youtube," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 91–118, 2010.

[7] L. Plissonneau, T. En-Najjary, and G. Urvoy-Keller, "Revisiting web traffic from a dsl provider perspective : the case of youtube," in *Proc. of the 19th ITC Specialist Seminar*, 2008.

[8] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[9] Builtwith.com, "Dailymotion video website categories," http://trends.builtwith.com/media/DailyMotion-Video.