HAL
open science

Géodistribution des tags et des vues dans Youtube

Stéphane Delbruel, François Taïani

To cite this version:

Stéphane Delbruel, François Taïani. Géodistribution des tags et des vues dans Youtube. Conférence d'informatique en Parallélisme, Architecture et Système. Compas'2015., Jun 2015, Lille, France. hal-01162568

HAL Id: hal-01162568

https://inria.hal.science/hal-01162568

Submitted on 10 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Compas'2015 : Parallélisme/ Architecture / Système
Lille, France, du 30 juin au 3 juillet 2015

Géodistribution des tags et des vues dans Youtube

Stéphane Delbruel[1], François Taïani[2]

[1]Université de Rennes 1 - IRISA - Inria
[2]Université de Rennes 1 - IRISA - ESIR - Inria
{stephane.delbruel,francois.taiani}@irisa.fr

Résumé
Dans cet article, nous analysons la corrélation entre la distribution géographique des vues d'une vidéo et les tags de cette vidéo au sein d'un dataset YouTube. Nous montrons que les tags peuvent servir d'indice sur la diffusion géographique d'une vidéo, avec certains tags très fortement liés à des zones géographiques bien définies. Cette corrélation peut être exploitée pour prédire correctement un minimum de 68% des vues pour une majorité des vidéos.

Mots-clés : vidéos en ligne, tags, distribution géographique, système de distribution de contenu

---

1. Introduction

La diffusion de vidéos en ligne représente aujourd'hui une des plus grandes sources de trafic mondial, avec des rapports lui attribuant jusqu'à 60% du trafic d'un FAI en pic de consommation [6]. Une grande part de ce trafic est causée par des Contenus Générés par les Utilisateurs (*User Generated Content* ou *UGC*) tels que Youtube, Dailymotion ou Vimeo : Youtube, par exemple, compte pour 18.69% de l'ensemble du trafic en Amérique du Nord, 28.73% en Europe et jusqu'à 31.22% en Asie [1]. Stocker, traiter et servir cette quantité de données est un défi constant d'ingénierie pour à la fois les services UGC et les FAI.
Mieux comprendre, et éventuellement pouvoir prédire, quelle vidéos sont vues dans quelles zones géographiques peut permettre d'améliorer substantiellement les infrastructures de distribution des services UGC. Les tags associées aux vidéos en ligne peuvent justement apporter cette information, mais cet aspect des services UGC a jusqu'à maintenant été très peu étudié. C'est ce que nous nous proposons d'aborder dans ce travail.
Dans ce papier, nous présentons et analysons un dataset Youtube et étudions la relation entre les tags d'une vidéo et où cette vidéo est vue, en utilisant les tags comme marqueurs géographiques des tendances de chaque pays. En s'appuyant sur nos découvertes, nous explorons la possibilité de prédire les vues d'une vidéo en fonction de ses tags. Nous montrons que même avec une approche simple de prédiction, nous sommes capables de prédire au minimum 68% des vues d'une vidéo pour une majorité de vidéos.

2. Problématique

Comprendre comment les vidéos d'un service en ligne sont consommées est primordial pour améliorer les infrastructures qui supportent ce service. De précédents travaux ont mis l'accent

sur la popularité et la temporalité de l'évolution de la distribution des contenus UGC [3]. Certains de ces travaux ont par exemple mis en lumière le potentiel des systèmes VoD (*Video on Demand*) assistés par les pairs [11, 7] pour faire face à la distribution très étalées (*long tail distributions*) de la popularité des vidéos dans les services UGC. D'autres ont cherché à quel point le graphe des vidéos recommandées dans Youtube affecte le comportement des usagers [12], et suggéré l'utilisation d'architectures pair-à-pair pour exploiter ce lien [8, 4].
Bien que particulièrement utiles, la plupart de ces travaux assument que la demande des systèmes vidéos UGC est uniformément distribuée ou presque d'un point de vue géographique. Or, il s'avère que la distribution géographique est primordiale dans la manière dont les vidéos des systèmes UGC sont partagées et consommées. Plus spécifiquement dans Youtube, de précédentes recherches ont montré qu'une majorité de vidéos ont un ensemble de vues fortement localisé [2]. Cette forte localité de la demande a des conséquences majeures sur la conception des infrastructures des systèmes vidéos UGC, telles que les politiques de cache dans les CDN, ou les stratégies de routage et de réplication dans les systèmes aidés par les pairs. Malheureusement, en dehors de quelques exemples notables [10, 9], peu de travaux jusqu'ici ont étudié comment prédire la localité d'une vidéo, et aucun n'a, à notre connaissance, considéré le rôle des tags attachés par les utilisateurs aux vidéos d'un service en ligne.

3. Tags, Vues, et géo-distribution dans Youtube

3.1. Le dataset

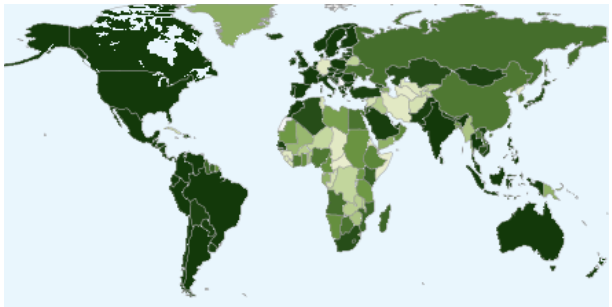

FIGURE 1 – Carte de popularité de la vidéo la plus vue de notre dataset *Justin Bieber - Baby ft. Ludacris*, telle que fournie par Youtube

Nous utilisons un dataset collecté par notre groupe de recherche en mars 2011 [8]. Le point de départ de ce dataset sont les 10 vidéos les plus populaires dans 25 pays différents, obtenues via l'API publique de Youtube. Le dataset a ensuite été collecté par échantillonnage "boule de neige" via le graphe des vidéos recommandées par Youtube. Pour chaque vidéo récoltée, le dataset contient entre-autres : l'identifiant de la vidéo, son titre, son nombre de vues, sa carte de popularité [1], et un ensemble de tags associés à cette vidéo. La carte de popularité d'une vidéo indique pour chacun des 235 pays présents dans le standard ISO 3166-1-alpha-2, via un entier de 0 à 61, quelle est la popularité de cette vidéo dans ce pays. Par exemple la Figure 1 montre la carte de popularité de la vidéo la plus vue de notre dataset (*Justin Bieber - Baby ft. Ludacris*) au travers d'une carte choroplèthe. Le dataset dans sa forme brute contient 1.063.844 vidéos uniques, mais toutes n'ont pas un ensemble complet de méta-données. Pour l'analyse présentée dans ce papier, nous évinçons les vidéos qui ne contiennent pas de tags (6.736 vidéos), ou avec des geodata incorrectes, pour ne retenir que 691.349 vidéos.

3.2. Notes et Mesures

Pour des raisons de clarté, nous utilisons la notation suivante dans le reste du document : $\mathcal{V}$ est l'ensemble des vidéos dans notre dataset nettoyé. Pour chaque vidéo $v \in \mathcal{V}$ nous utilisons les

1. Cette information n'est désormais plus disponible. YouTube a changé leur API depuis Septembre 2013 et interdit l'accès à la distribution géographique des vues d'une vidéo.

informations suivantes :

— $\mathtt{tags}(v)$ est l'ensemble des tags de cette vidéo. Par exemple, la vidéo la plus vue de notre dataset (Figure 1) est associée avec les tags *Justin*, *Bieber*, *Island*, *Def*, *Jam* and *Pop*.
— $\mathit{tot_views}(v)$ est le nombre total de vues d'une vidéo ;
— $\mathbf{pop}(v)$ est le vecteur de la popularité d'une vidéo donné par Youtube. $\mathbf{pop}(v)[c]$ est l'entier représentant la popularité de v dans le pays c.
— $\mathbf{views}(v)$ est le vecteur des vues de la vidéo dans chaque pays du monde. On note $\mathbf{views}(v)[c]$ le nombre de fois où v a été vue dans un pays c. Cette information n'est pas directement fournie par Youtube.

A partir de ces informations, nous calculons pour chaque tag t les ensembles et statistiques suivantes :

— $\mathtt{vidéos}(t)$ est l'ensemble des vidéos contenant t dans leurs tags.

$$\mathtt{videos}(t) = \left\{v \in \mathcal{V} \mid t \in \mathtt{tags}(v)\right\} = \mathtt{tags}^{-1}(t)$$

— $\mathtt{freq}(t)$ est le nombre d'occurrences de t : $\mathtt{freq}(t) = |\mathtt{videos}(t)|$
— $\mathit{tot_views}(t)$ est le nombre total de vues associées à t, i.e. l'agrégation des vues des vidéos contenant t.

$$\mathit{tot_views}(t) = \sum_{v \in \mathtt{videos}(t)} \mathit{tot_views}(v)$$

La signification exacte de $\mathbf{pop}(v)$ n'est pas documentée par YouTube. Il est peu probable que l'entier soit directement proportionnel au nombre de vues de la vidéo v par pays $\mathbf{views}(v)$. Appliquée à la vidéo la plus vue de notre dataset (*Justin Bieber - Baby ft. Ludacris*, figure 1), cela signifierai que cette vidéo a été autant vue aux USA (population 318,5M) qu'en Islande (population 329.040). Nous considérons plutôt le vecteur $\mathbf{pop}(v)$ comme *l'intensité* de la vidéo dans chaque pays, i.e. un nombre proportionnel aux part des vues de YouTube que représente chaque pays avec :

$$\mathbf{pop}(v)[c] = \frac{\mathbf{views}(v)[c]}{\mathbf{ytube}[c]} \times K(v) \tag{1}$$

où $\mathbf{views}(v)[c]$ est le nombre de vues de v dans le pays c, $\mathbf{ytube}[c]$ est le nombre total de vues YouTube dans le pays c, et $K(v)$ est un facteur de normalisation, dépendant de chaque vidéo, pour mettre à l'échelle des valeurs entre $[0 - 61]$.

Ni $\mathbf{ytube}[c]$ ni $K(v)$ nous sont fournis. Pour les estimer, nous utilisons la distribution du trafic de YouTube donné par Alexa Internet Inc. [2], pour approximer la part des vues YouTube par pays :

$$\mathbf{ytube}[c] = \mathbf{p}_{yt}[c] \times T_{yt} \simeq \widehat{\mathbf{p}}_{yt}[c] \times T_{yt} \tag{2}$$

où $\mathbf{p}_{yt}[c]$ est la part des vues YouTube dans un pays c, T_{yt} est le nombre de total de vues YouTube, et $\widehat{\mathbf{p}}_{yt}[c]$ est le trafic YouTube estimé par Alexa pour un pays c.

En utilisant les autres données du dataset, il est possible d'éliminer $\mathbf{ytube}[c]$, $K(v)$ et T_{yt} de (1), et ainsi d'estimer la distribution géographique des vues d'une vidéo $\mathbf{views}(v)[c]$. Pour chaque tag t, nous dérivons le nombre de vues associées à t dans le pays c (noté $\mathbf{views}(t)[c]$), i.e. le nombre agrégé des vues dans un pays c des vidéos contenant t comme tag.

$$\mathbf{views}(t)[c] = \sum_{v \in \mathit{videos}(t)} \mathbf{views}(v)[c]$$

2. http://www.alexa.com/

TABLE 1 – Les 5 tags les plus fréquents

tag	#occur	#vues	moyenne #vues
the	30686	13.157.705.562	428.785
video	27239	12.898.383.171	473.526
music	23128	12.640.171.764	546.531
2010	22014	3.349.620.292	152.158
funny	21645	13.550.709.569	626.043

TABLE 2 – les 5 tags les plus vus (global)

tag	#occur	#vues	moyenne #vues
funny	21645	13.550.709.569	626.043
pop	7877	13.318.507.233	1.690.809
the	30686	13.157.705.562	428.785
video	27239	12.898.383.171	473.526
music	23128	12.640.171.764	546.531

Dans cette analyse, nous sommes également particulièrement intéressés par capturer la diffusion géographique d'un tag (resp. sa concentration), et en contrastant cette diffusion avec les vidéos associées à ce tag. Pour y parvenir, nous utilisons l'entropie de Shannon $H(t)$ sur la distribution des vues d'un tag t (resp. video v) pour chaque pays.

$$H(x) = -\sum_{c \in World} \mathbf{p}_{\text{geo}}(x)[c] \times \log_2\left(\mathbf{p}_{\text{geo}}(x)[c]\right)$$

où x est soit une vidéo soit un tag, et $p_{\text{geo}}(x)[c]$ représente la proportion des vues de cette vidéo ou de ce tag pour un pays c :

$$\mathbf{p}_{\text{geo}}(x)[c] = \frac{\mathbf{views}(x)[c]}{tot_views(x)}$$

Une entropie élevée signifie que le tag (ou la vidéo) tend à être diffusée de manière uniforme parmi les différents pays. Au contraire, une entropie basse dénote un tag (ou une vidéo) dont les vues ont tendances à être concentrées dans très peu de pays. Par exemple, la vidéo avec le plus grand nombre de vues de notre dataset, *Justin Bieber - Baby ft. Ludacris* montrée en Figure 1, possède une entropie de $5,06$. Cette valeur est proche du maximum possible étant de $\log_2(235) = 7,87$, qui correspondrai à une vidéo également distribuée parmi les 235 pays suivis par YouTube. En contraste, la plus basse valeur possible est $\log_2(1) = 0$, correspondant à un tag ou une vidéo dont les vues sont toutes dans un seul pays.

3.3. Tag et distribution des vues

Notre dataset contient 7.717.815 occurrences de tag, amenant à un nombre moyen de 11,18 tags par vidéo. Ces occurrences de tags sont composées de 705.415 tags uniques, un nombre élevé en accord avec de précédents travaux [5]. Ce grand nombre de tags uniques peut être expliqué par la présence de tags composés (e.g. "*korean pop*" est différent de "*korean*" "*pop*", qui compte pour deux tags), ou les fautes de frappes ("*(music*" ou "*music␣*" au lieu de "*music*"), et l'utilisation de plusieurs langues.

Un échantillon des 5 tags les plus fréquents est présenté en Table 1, et les 5 plus vus en Table 2. Ces deux tables mettent en valeur quelques précieux éléments de l'usage des tags dans YouTube : bien que quelques mots de grammaire soient présents (*the*), la plupart traitent de contenu (*video*, *funny*). Les éléments grammaticaux peuvent être expliqués par l'ancienne utilisation des espaces dans Youtube, pour séparer les tags (aujourd'hui, des virgules), qui amène des tags voulus composés tel que *the␣rock* à être considérés comme deux tags. les tags les plus vus ne sont pas forcément les plus fréquents : c'est particulièrement vrai pour *pop*, second tag le plus vu (Table 1), avec seulement 7877 occurrences. Les vidéos correspondantes sont issues très majoritairement de la catégorie "Music", avec un nombre moyen de vues par vidéo élevé (1.690.809 vues, 2,7 fois plus que les vidéos contenant le tag *funny* par exemple). La même observation s'applique aux tags associés tel que *hip*, et *records*.

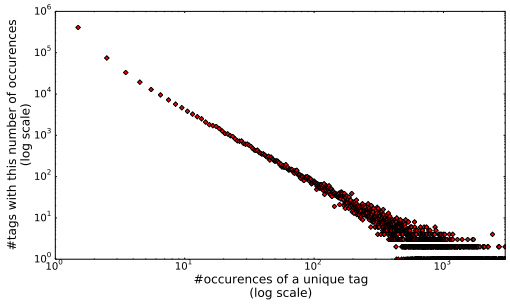

FIGURE 2 – La distribution des tags suit une loi de puissance ($y = K \times x^{-\alpha}$) comme souvent observé en folksonomie et langages naturels

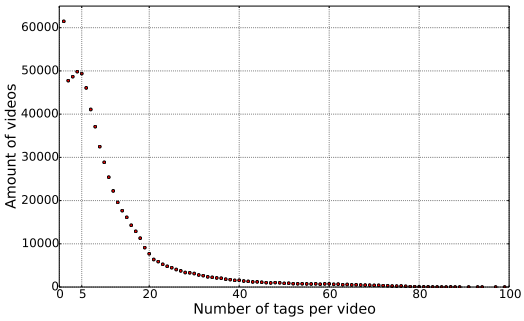

FIGURE 3 – Les tags sont largement utilisés pour décrire une vidéo, avec 50% des vidéos, comportant au moins 11 tags

TABLE 3 – Top 5 des pays par vues pour *pop*

pays	#vues	% total vues
United-States	4.700.159.350	35,2%
United-Kingdom	759.449.112	5,7%
Brazil	751.342.295	5,6%
Mexico	603.876.310	4,5%
India	586.339.771	4,4%

TABLE 4 – Top 5 des pays pour *bollywood*

pays	#vues	% total vues
India	200.956.055	39,8%
United-States	124.461.447	24,7%
United-Kingdom	29.506.586	5,8%
Pakistan	25.218.518	5,0%
Germany	12.842.983	2,5%

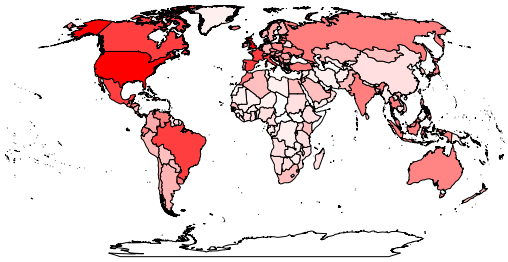

FIGURE 4 – Distribution du tag *pop*

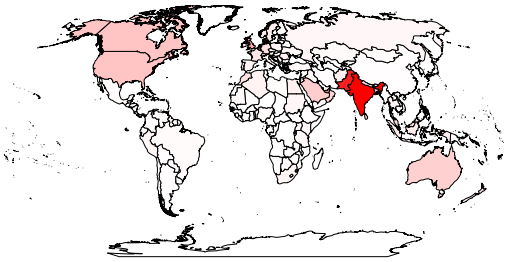

FIGURE 5 – Distribution du tag *bollywood*

La distribution fréquentielle des tags (Figure 2) démontre une loi de puissance typique, qui se retrouve communément dans les langages naturels et folksonomies. En particulier, 462.549 tags (66%) n'apparaissent qu'une fois.

Les descriptions par tag d'une vidéo donnée sont relativement riches (Figure 3), avec une moyenne de 11,18 tags par vidéo comme mentionné plus haut. Une raison pour laquelle les vidéos comportent un nombre de tags raisonnable peut-être que les utilisateurs ont tendance à taguer leurs vidéos pour attirer plus de vues.

Par la suite, et pour éviter les artefacts causés par les vidéos avec un nombre trop faible de vues, nous prendront en compte seulement celles avec plus de 1000 vues. Nous limiterons également notre discussion aux tags iso-latin1 (91,03% de toutes les occurrences de tags).

3.4. Distribution géographique des tags

En termes de distribution géographique, nous pouvons observer des distributions géographiques grandement différentes d'un tag à un autre, avec certains tags largement répandus, et d'autres concentrés dans seulement quelques pays. Nous discuterons plus en détail sur ces

TABLE 5 – Les 5 tags avec le moins (g.) et le plus (d.) d'entropie (pour #occurs > 100)

tag	H(t)	#occurs	nb. vues	moyenne des vues
piologo	0,04	101	3.985.341	39.458
mundo canibal	0,06	134	4.147.868	30.954
kvarteret	0,10	102	7.313.481	71.700
skatan	0,11	106	7.741.235	73.030
partoba	0,18	272	7.183.083	26.408

tag	H(t)	#occurs	nb. vues	moyenne des vues
recovery	4,90	230	557.869.571	2.425.519
dominic	4,87	103	338.555.257	3.286.944
fifa	4,83	2722	690.091.145	253.524
passat	4,79	142	41.809.394	294.432
afraid	4,78	131	244.659.961	1.867.633

deux cas, en se penchant sur les deux tags *pop*, et *bollywood*. Le top 5 des pays consommateurs de ces tags sont indiqués en tables 3 et 4, et la distribution de leurs vues est représentée dans les figures 4 et 5. Sur ces cartes choroplèthes, une saturation plus élevée de rouge représente une proportion plus élevée de vues dans le pays.
Les vues associées au tag *pop* (entropie 4,25) tendent à être très distribuées à travers le monde (Table 3) et Figure 4). Le pays avec le plus de vues associées étant les USA, ils ne représentent pourtant que 35,2% du nombre de vues global du tag *pop*.
A contrario, le tag *bollywood* (Table 4 et Figure 5) est beaucoup plus concentré dans peu de pays, ce qui se reflète sur son score d'entropie : 3,24. Les vues pour *bollywood* sont principalement en Inde et au Etats-unis (64,5%), comme attendu pour des raisons culturelles et linguistiques, avec trois pays supplémentaires comptant pour un autre 11,3% du total de vues. En mettant en cache, ou en plaçant de manière pro-active des copies des vidéos contenant *bollywood* dans ces 5 pays, un service vidéo de type UGC couvrirait 75,8% des vues de ce tag, représentant une part substantielle du trafic.

3.5. Analyse entropique

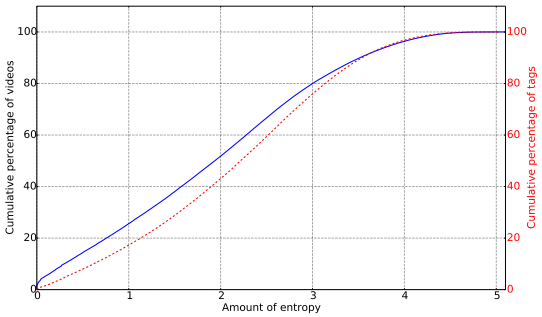

FIGURE 6 – CDF des vidéos (ligne) et tags (pointillés) en fonction de l'entropie

Afin de comprendre plus profondément comment les tags YouTube sont distribués, nous nous tournons vers l'entropie. L'entropie d'un tag ou d'une vidéo procure une idée générale de la diffusion géographique de ce tag ou cette vidéo, comme illustré par les deux tags que nous venons d'analyser. La Figure 6 montre la CDF de l'entropie des vidéos (ligne) et des tags (pointillés) dans notre dataset. Les deux courbes sont similaires : les valeurs d'entropie paraissent uniformément réparties pour des valeurs allant jusqu'à 3 (ce qui correspond environ à 80% de l'ensemble des tags et vidéos). Seulement 2,81% de l'ensemble des vidéos ont une entropie supérieure à 4. Ces nombres soulignent qu'une part substantielle de vidéos ont une faible entropie (i.e. dont les vues sont géographiquement concentrées) : 40% de l'ensemble des vidéos ont une entropie inférieure à $1,5$. Comme point de référence, cela est légèrement en dessous de la valeur qu'une vidéo obtiendrait si elle était uniformément distribuée parmi seulement 3 pays ($\log_2(3) = 1.585$).
La Figure 7 met en couleur la relation entre l'entropie d'une vidéo et son nombre de vues. Comme attendu et comme introduit dans de précédents travaux [8], les vidéos populaires, en particulier celles au delà de 10^6 vues tendent à avoir une entropie élevée, signifiant que leurs vues sont largement distribuées. C'est aussi vrai pour les vidéos avec peu de vues, avec une

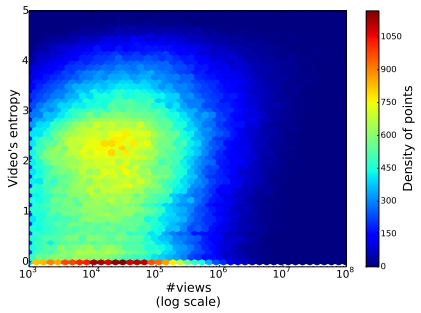

FIGURE 7 – Entropie des vidéos en fonction de leur nb. de vues

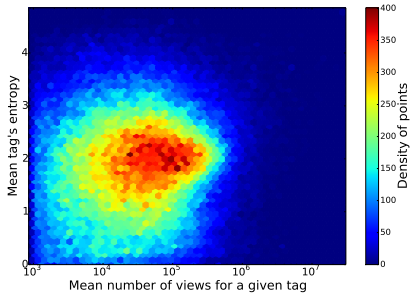

FIGURE 8 – Entropie des tags en fonction de leur nb. de vues

concentration des vidéos dont les vues sont dans l'intervalle $[10000, 200000]$ et dont l'entropie est proche de $2,5$. En dehors de ces zones à concentration élevée et pour des vidéos de moins de 10^6 vues, les valeurs d'entropie tendent à être équitablement distribuées, avec deux points secondaires de concentration autour des valeurs $1,5$ et 0. Ces distributions montrent que les vidéos grandement populaires ont besoin en moyenne d'être accédées de partout dans le monde, étant donné que leur entropie est grande. En revanche, les vidéos les moins vues ont une entropie relativement basse et bénéficieraient grandement d'une prédiction précise de leur distribution géographique. Ce que nous avançons dans ce papier, c'est que cette information peut être extraite des tags, tout du moins en partie.

En se penchant sur les tags, la Figure 8 montre la relation entre le nombre moyen de vues d'un tag et sa valeur entropique moyenne (les deux moyennes étant calculées sur les vidéos où apparaît le tag), représentée sur un graphe de densité. Comme pour les vidéos, la plus grande concentration de tags se situe pour des valeurs d'entropie autour de 2, et une moyenne des vues à 100.000, mais la diffusion des tags dans le reste du graphe demeure substantielle, i.e. dans la zone avec une entropie < 3 et un nombre moyen de vues compris entre 10.000 et 200.000.

La présence d'un nombre important de vidéos et de tags ayant une faible entropie, tout en représentant un nombre non–négligeable de vues, nous mène à vouloir étudier comment la distribution géographique des vues d'une vidéo peut être prédite en étudiant simplement les tags qui lui sont associés.

4. Prédire les vues à partir des tags

L'analyse de la section précédente a démontré que la distribution géographique des tags et celle des vidéos sont fortement corrélées. Dans cette section, nous irons plus loin en explorant plus en détail le potentiel prédictif des tags, en terme de distribution géographique de la consommation, pour leurs vidéos associées. Nous utiliserons pour cela une technique de prédiction basique, et l'évaluerons par *validation croisée* (Section 4.2), où nous enlevons un pourcentage des vidéos (notées $\mathcal{V}_{\text{test}}$) afin de prédire leur distribution, sur la base des vidéos restantes (notées $\mathcal{V}_{\text{train}}$).

4.1. Approche prédictive

Pour une vidéo $v \in \mathcal{V}_{\text{test}}$ associée à l'ensemble des tags $\mathtt{tags}(v)$, nous prédisons la distribution géographique des vues de v $\widehat{\mathbf{p}_{\text{geo}}}(v)$ par la moyenne de la distribution géographique des tags de v :

$$\widehat{\mathbf{p}_{\text{geo}}}(v) = \mathop{\mathbb{E}}_{t \in \mathtt{tags}(v)} \left(\mathbf{p}_{\text{geo}}^{\mathcal{V}_{\text{train}}}(t) \right)$$

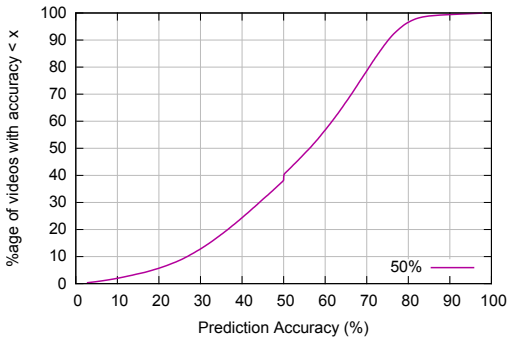

FIGURE 9 – Précision des prédictions obtenues (distribution cumulative, 50% de retenu sur $\mathcal{V}_{[10k,200k]}$)

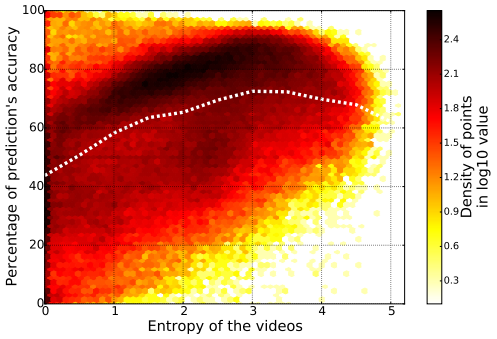

FIGURE 10 – Précision des prédictions en fonction de l'entropie des vidéos (50% de retenu sur $\mathcal{V}_{[10k,200k]}$. La moyenne en pointillés.

où $\mathbf{p}_{geo}^{\mathcal{V}_{train}}(t)$ est le vecteur de la distribution géographique du tag t dans le dataset $\mathcal{V}_{train}$ (qui ne prends pas en compte les vidéos de $\mathcal{V}_{test}$). Notre but est que $\widehat{\mathbf{p}_{geo}}(v)$ soit aussi près que possible de $\mathbf{p}_{geo}(v)$, l'actuel vecteur de distribution géographique de v.

4.2. Validation croisée

Pour évaluer cette méthode de prédiction, nous nous concentrons sur les vidéos ayant entre 10.000 et 200.000 vues, notées $\mathcal{V}_{[10k,200k]}$, 287.700 videos. Cette classe de vidéo contient à la fois des vidéos concentrées dans peu de pays (faible entropie), et à la fois des vidéos distribuées mondialement, comme aperçu dans la Figure 7. Nous utiliserons le reste du dataset pour servir d'ensemble de référence à la prédiction de la distribution géographique de ces vidéos.
Pour construire notre ensemble de référence, nous enlevons de notre dataset complet $\mathcal{V}$ la moitié des vidéos dans $\mathcal{V}_{[10k,200k]}$. On nomme les vidéos retirées $\mathcal{V}_{[10k,200k]}^{-0.5}$. Ce procédé défini nos ensemble de référence et de test :

$$\begin{aligned}\mathcal{V}_{test}^{-0.5} &= \mathcal{V}_{[10k,200k]}^{-0.5}\\ \mathcal{V}_{train}^{-0.5} &= \mathcal{V} \setminus \mathcal{V}_{[10k,200k]}^{-0.5}\end{aligned}$$

On tente ensuite de prédire la distribution géographique des vidéos de $\mathcal{V}_{[10k,200k]}^{-0.5}$ à partir des informations contenues dans $\mathcal{V}_{train}^{-0.5}$.

4.2.1. Qualité des prédictions

Afin de mesurer la divergence entre notre prédiction $\widehat{\mathbf{p}_{geo}}(v)$ et la distribution géographique originelle d'une vidéo $\mathbf{p}_{geo}(v)$ nous calculons la proportion de vues mal positionnées par la prédiction, $\mathbf{p}_{wrong}(v)$:

$$\mathbf{p}_{wrong}(v) = \frac{1}{2} \times \sum_{c \in World} \left| \mathbf{p}_{geo}(v)[c] - \mathbf{p}_{geo}(v) \right|$$

La somme est ensuite divisée par deux pour éviter de compter deux fois les erreurs (pour les pays qui en manquent, et ceux qui en ont en trop). Nous définissons notre métrique finale, la *précision* de la prédiction, comme le complément de ces vues mal placées :

$$\mathbf{p}_{accurate}(v) = 1 - \mathbf{p}_{wrong}(v)$$

Une précision de 1 signifie que la prédiction est parfaite face à l'original, 0 indiquant qu'il n'y a aucun pays commun entre la prédiction et l'original.
La Figure 9 montre la précision de nos prédictions lorsque 50% des vidéos sont retirées de $\mathcal{V}_{[10k,200k]}$. Nous obtenons une précision moyenne de 64,2%, et une précision médiane de 68,1%. Cela établit que notre approche est capable de prédire avec un minimum de 68% de précision la distribution des vues pour une majorité de vidéos, un résultat particulièrement encourageant étant donné la simplicité de notre technique.
En termes d'entropie (Figure 10), on note que les vidéos avec une entropie importante (> 3) sont prédites avec une précision moyenne proche de 70%. Bien que les vidéos avec une faible entropie (< 2) présentent des résultats plus faibles, avec une précision de prédiction moyenne de 55%, la figure révèle que certaines de ces vidéos s'y prêtent très bien. En prenant en compte le fait que notre approche très basique ne fait pas de distinction entre les tags et n'effectue aucune forme de régression, cela démontre le fort potentiel des tags pour prédire la distribution géographique, même pour des vidéos qui sont concentrées dans très peu de pays.
Ces résultats confirment notre hypothèse originale : Les tags peuvent nous permettre de prédire de manière précise la distribution géographique des vidéos UGC, avec des résultats encourageants obtenus pour les vidéos avec un nombre relativement réduit de vues (de 10.000 à 200.000 vues).

5. Conclusion

Dans ce papier nous avons proposé une analyse de la distribution géographique des tags dans YouTube en utilisant un dataset composé de 691.349 vidéos, associées à 7.717.815 occurrences de tag pour 705.415 tags uniques, et totalisant 173.288.616.473 vues. Notre analyse démontre que tags et vidéos démontrent un large spectre de dissémination à travers le monde, avec certains tags concentrés dans très peu de pays (faible entropie) et d'autres plus équitablement répandus (grande entropie). Nous avons également montré que si les vidéos très populaires ont tendances à être des "vidéos globales", ce n'est pas vrai des vidéos vues entre 10.000 et 200.000 fois, qui elles ont un comportement beaucoup plus divers. En poussant notre analyse plus en avant, nous avons démontré un lien entre la diffusion géographique des tags et des vidéos qui leur sont reliés, et ainsi permettre, à l'aide de techniques très simples, de pouvoir prédire de manière conséquente la distribution géographique des vues des vidéos en se basant uniquement sur ses tags. Nos résultats (un minimum de 68% des vues prédites de manière précise pour la majorité des vidéos) démontrent un fort potentiel des tags à la précision du placement et du caching, en particulier couplé avec des techniques d'apprentissage. Nous pensons que ce travail ouvre de grandes perspectives pour améliorer l'implémentation de systèmes géo-répliqués de stockage à large échelle.

Remerciements

Ce travail à reçu un soutien du gouvernement français attribué au laboratoire d'excellence CominLabs (Project "DeSceNt : Plug-based Decentralized Social Network") et géré par l'ANR dans le programme "Investing for the future" sous la référence Nb. ANR-10-LABX-07-01. Ce travail est également porté par une bourse de l'Université de Rennes 1 au sein de son programme "Politique doctorale 2013" pour le projet *Towards a Decentralized Embryomorphic Storage System*.

Bibliographie

1. *Global Internet Phenomena Report : 2H 2013*. – Rapport technique, Sandvine Incorporated, 2013.
2. Brodersen (A.), Scellato (S.) et Wattenhofer (M.). – YouTube around the world : Geographic popularity of videos. – In *WWW*, 2012.
3. Cha (M.), Kwak (H.), Rodriguez (P.), Ahn (Y.-Y.) et Moon (S.). – I tube, you tube, everybody tubes : Analyzing the world's largest user generated content video system. – In *IMC*, 2007.
4. Cheng (X.) et Liu (J.). – NetTube : Exploring social networks for peer-to-peer short video sharing. – In *INFOCOM*, 2009.
5. Geisler (G.) et Burns (S.). – Tagging video : conventions and strategies of the youtube community. – In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 480–480. ACM, 2007.
6. Guillemin (F.), Kauffmann (B.), Moteau (S.) et Simonian (A.). – Experimental analysis of caching efficiency for youtube traffic in an isp network. – In *International Teletraffic Congress*, pp. 1–9. IEEE, 2013.
7. Huang (Y.), Chen (Y.-F.), Jana (R.), Jiang (H.), Rabinovich (M.), Reibman (A.), Wei (B.) et Xiao (Z.). – Capacity analysis of mediagrid : a p2p iptv platform for fiber to the node (fttn) networks. *IEEE Journal on Selected Areas in Communications*, vol. 25, n1, 2007, pp. 131–139.
8. Huguenin (K.), Kermarrec (A.-M.), Kloudas (K.) et Taïani (F.). – Content and geographical locality in user-generated content sharing systems. – In *Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video*, *NOSSDAV '12*, NOSSDAV '12, pp. 77–82. ACM, 2012.
9. Sastry (N.), Yoneki (E.) et Crowcroft (J.). – Buzztraq : Predicting geographical access patterns of social cascades using social networks. – In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pp. 39–45. ACM, 2009.
10. Scellato (S.), Mascolo (C.), Musolesi (M.) et Crowcroft (J.). – Track globally, deliver locally : Improving content delivery networks by tracking geographic social cascades. – In *WWW*, 2011.
11. Yin (H.), Liu (X.), Zhan (T.), Sekar (V.), Qiu (F.), Lin (C.), Zhang (H.) et Li (B.). – LiveSky : Enhancing CDN with P2P. *ACM TOMCCAP*, vol. 6, 2010, pp. 16 :1–16 :19.
12. Zhou (R.), Khemmarat (S.) et Gao (L.). – The impact of youtube recommendation system on video views. – In *IMC*, 2010.