Projet Data Mining

Pierre Turpin, Jean-Marie Comets

25 mars 2014

Table des matières

Au vu des nombreux problèmes de "scaling" que nous pouvions rencontrer avec le jeu de données prévu, l'intégralité de ce rapport repose sur l'analyse d'un échantillon de données. Bien entendu, avant d'étudier les données, nous avons mélangé le jeu de données et pris un échantillon fixe pour l'ensemble de l'étude.

1 Caractérisation du flux vidéo

1.1 Popularité

En étudiant les différents indicateurs quantitatifs (attributs *_count), représentant le nombre de vues d'un flux, nous avons pu remarquer que quasiment toutes celles-ci sont indépendantes, mis à part l'attribut *stream_ count*, qui ne représente que la somme du *embedded_ count* et du *site_ count*.

Nous avons donc retiré cet attribut, pour pouvoir définir la notion de **popularité** d'un flux, correspondant à une somme normalisée des différents indicateurs.

Cet indicateur nous sert d'heuristique pour établir les différentes catégorisations suivantes :

1.2 Catégorisation

Un seul attribut permettant de catégoriser les différents flux est disponible, et ce uniquement à partir du jeu de données XML : *subcategory*. En remarquant que cet attribut concerne à la fois la catégorie du jeu et sa plateforme, nous avons séparé ces deux informations.

Ainsi, dans la figure 1, nous n'observons pas la plateforme PC, qui devrait cependant regrouper beaucoup de flux (plateforme non définie là où le type de jeu est bien défini). Cependant, nous pouvons remarquer qu'entre les différentes consoles de jeu, c'est la plateforme XBOX qui a le plus de succès.

Nous remarquons à partir de la figure 2 donc que la catégorie **strategy** se détache des autres, prenant plus de 60% de la part du nombre de vues des flux. Ceci correspond bien à nos attentes, vu que Twitch est principalement connu pour des jeux développés pour PC, avec une préférence pour les jeux de stratégie/roleplay (Starcraft, League of Legends, etc...).

1.3 Les jeux vidéo populaires

Dans le jeu de données utilisé, la plupart des jeux étaient en doublons car ils n'étaient pas tous orthographiés de la même façon (majuscules/minuscules, espaces, ...). Nous n'avons pas pu, à cause de la taille des données, corriger toutes les entrées afin d'unifier l'écriture des jeux. Les résultats ne sont donc pas complètement exacts mais donnent

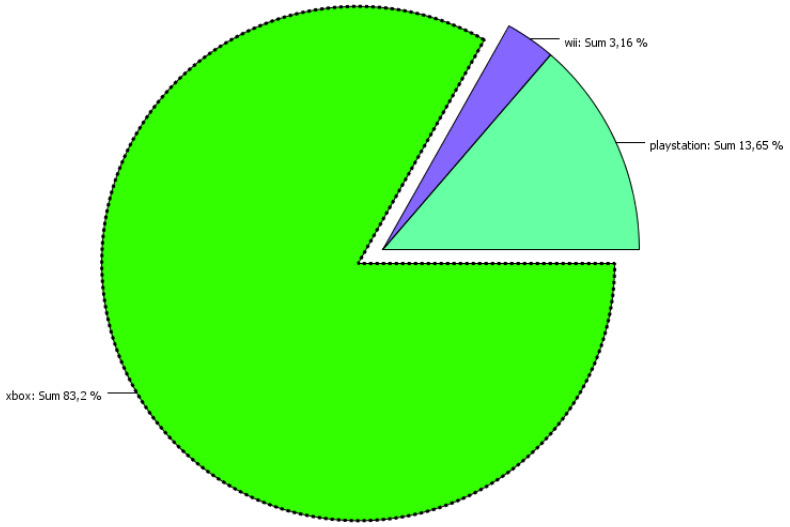

FIGURE 1: Parts des vues selon la plateforme (console) du jeu

tout de même une bonne approximation de la réalité.

En tout il y a 160 jeux différents en comptant les doublons. Nous avons établi la part de vue des différents jeux en groupant sur le champ *meta_ name* et en sommant l'heuristique de popularité. Comme une grande quantité de jeu était très minoritaire selon notre heuristique, nous avons regroupé ces derniers (en seuillant la popularité) dans une seule catégorie *misc*.

La figure 3 représente la popularité de chaque jeux. L'ensemble des jeux *misc* forment 19% de popularité tandis que 10 autres jeux prennent les 80% restant. Il y a donc une très grande disparité dans les jeux vidéo et une petite minorité de 10 jeux écrasent totalement 150 autres jeux.

TABLE 1: Classement des 10 jeux les plus populaires sur la plateforme Twitch.

Position	Jeux vidéo	Parts de popularité (en %)
1	StarCraft II	25.12
2	World of Warcraft : Cataclysm	11.25
3	Super Street Fighter IV	10.18
4	League of Legends	8.98
5	Gears of War	7.46
6	Battlefield 3	5.87
7	Call of Duty : Black Ops	3.58
8	The Ico & Shadow of Colossus Collection	3.26
9	Halo : Reach	3.23
10	Heroes of Newerth	2.21

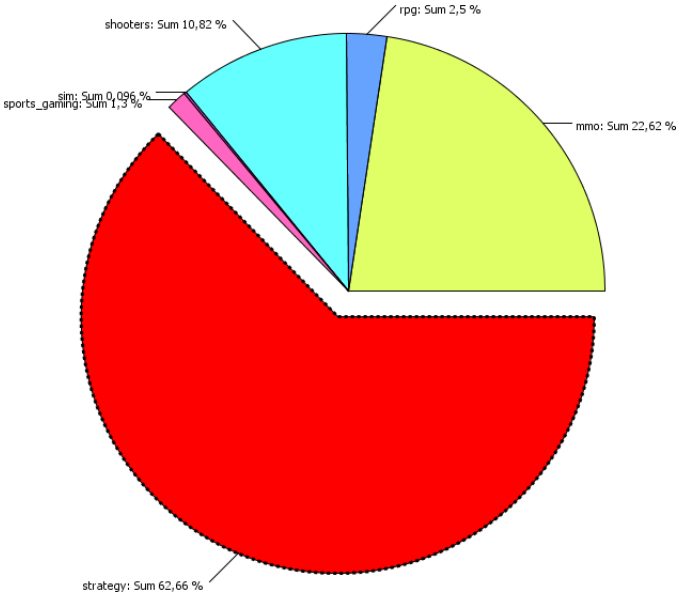

FIGURE 2: Parts des vues selon la catégorie du jeu

Le tableau 1 montre alors un classement des jeux les plus populaires sur la plateforme Twitch.

2 Prédiction de l'audience d'un flux

Un des intérêts majeur de cette étude est de pouvoir prédire l'audience d'un flux à partir d'informations simples sur le flux, telles que la localisation, la langue ou encore le type de jeu.

Nous nous sommes focalisés sur l'étude de la qualité d'un flux, ainsi que sa localisation et sa langue, pour pouvoir prévoir la popularité du flux.

2.1 Localisation/Langue

Avant de démarrer, la première impression en observant le jeu de données, a été le poids important des États-Unis dans l'audience de Twitch, vu que cette plateforme a été développée là-bas. Ça n'a donc pas été surprenant de voir nos différents calculs de clusters par localisation écrasés par le poids des États-Unis.

En ce qui concerne la langue, le résultat est encore plus flagrant, la langue anglaise est présente sur une majorité imposante des flux. Ce n'est donc pas surprenant de ne pas pouvoir prévoir quoi que ce soit à partir de cette information.

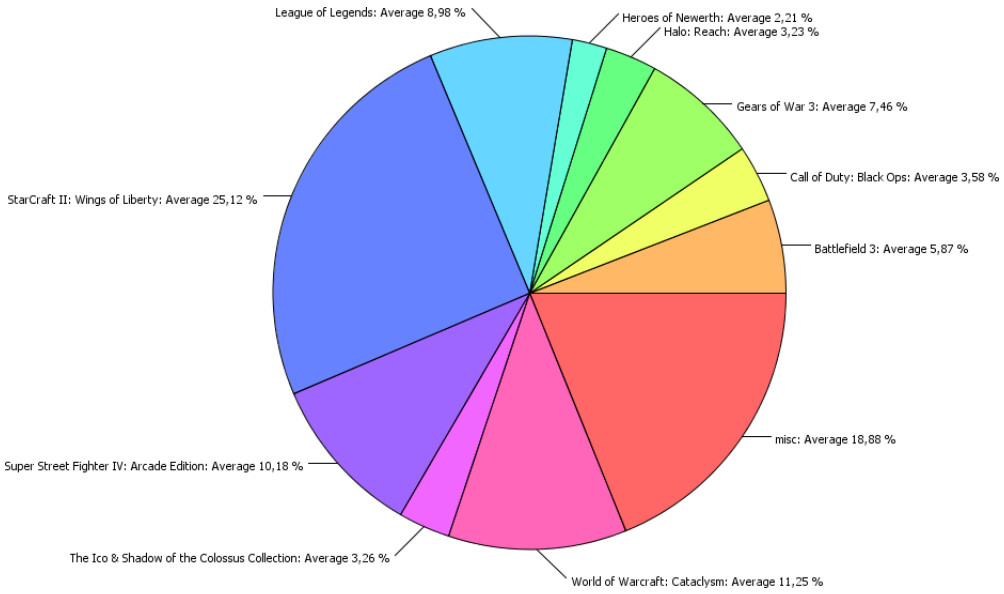

FIGURE 3: Parts de vues en moyenne des jeux vidéo

En conclusion, notre recherche de groupes ou de motifs récurrents à partir des informations de localisation a été infructueuse. Peut-être qu'à partir d'informations moins anonymes, avec par exemple la position géographique approximative du joueur, nous pensons qu'il serait possible de prédire l'audience du flux.

2.2 Qualité

Nous distinguons deux informations sur la qualité du flux : la qualité vidéo et la qualité audio. Au vu du faible nombre de valeurs discrètes, qui sont fortement distribuées par rapport au reste des attributs, nous avons décidé de ne pas étudier la qualité audio du flux.

Nous n'avons que quelques informations sur la qualité de la vidéo : les dimensions (*video_width* de la vidéo et *video_height*) et le taux d'émission du flux (*bitrate*). Nous avons donc multiplié ces différents attributs pour définir la qualité de la vidéo, à partir de laquelle nous pouvions essayer d'apprendre le succès prévu. Nous avons ignoré l'attribut *video_codec*, qui ne comporte que deux valeurs discrètes, dont une en très grande majorité.

Malheureusement nous avons pu confirmer que la qualité de la vidéo n'a aucune influence sur la popularité de la vidéo, vu que celle-ci est plutôt uniforme.

Remarque : nous avons préféré ne pas intégrer les graphiques représentant l'influence de la qualité, vu qu'ils ne mettent aucune information intéressante en valeur.

2.3 Mise en avant sur Twitch et exportation sur d'autres plateforme

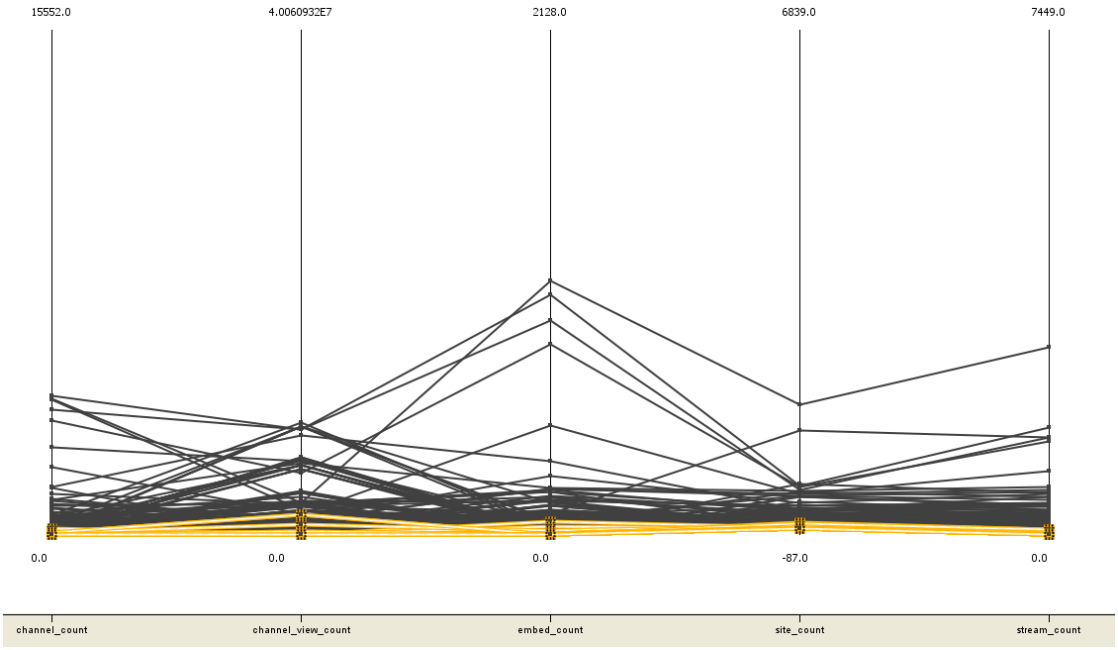

FIGURE 4: Vue en parallèle des compteurs de vues avec les flux non exportable surlignés.

Nous avons remarqué qu'il était possible d'activer l'exportation du flux vidéo sur une plateforme distante. Cela permet donc normalement d'étendre la visualisation et également le nombre de vue. Nous avons alors essayé de confirmer cette hypothèse.

Le champ *embed_ enabled* du jeu de données indique si le flux est exportable ou pas. Nous avons surligné et coloré les flux ne pouvant pas être exportés puis nous avons tracé leurs nombres de vues en comparant avec les flux exportables sur la figure 4. On voit alors clairement que les éléments colorés ont un nombre de vue largement moins important que les autres.

Cela confirme alors notre hypothèse et nous pouvons donc définir que l'activation de l'exportation d'un flux est une condition de succès et de popularité.

En parallèle à l'exportation des flux, Twitch met en avant certaines vidéos sur leur page principale. Logiquement, on peut émettre l'hypothèse que cela contribue fortement à la popularité.

Cette fois, nous avons fait la même chose qu'auparavant avec le champ *featured* indiquant si le flux est mis en avant ou pas. La figure 5 montre l'équivalant de précédemment, à savoir, le nombre de vues pour chaque flux avec un colorie sur ceux qui ne sont pas mis

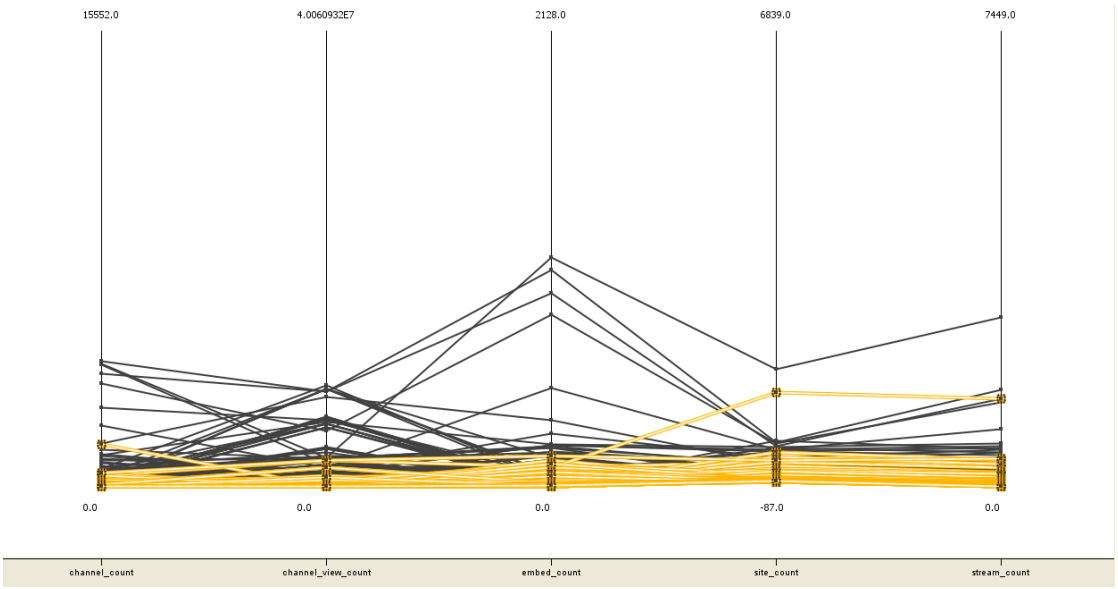

FIGURE 5: Vue en parallèle des compteurs de vues avec les flux non mis en avant surlignés.

en avant.

La encore on remarque qu'ils beaucoup moins vues que les autres. Cela confirme donc notre hypothèse. La mise en avant par la plateforme Twitch est donc une bonne condition de succès d'une vidéo.

Cependant, on remarque que ce résultat est moins marquant qu'avec l'exportation de la vidéo. L'exportation de la vidéo est donc plus important que la mise en avant par Twitch pour avoir beaucoup de vues sur sa vidéo.

3 Classement des "meilleurs" joueurs

Afin de trouver les utilisateurs ayant le plus de vues et donc le plus de popularité, nous avons regroupé les différentes vidéos appartenant au même utilisateurs en sommant le nombres de vues. En appliquant ensuite notre heuristique de popularité, on obtient alors la popularité totale de chaque utilisateurs. Nous avons ensuite trié cet ensemble afin de déterminé le classement de popularité des utilisateurs.

Pour vérifier que ce résultat est valide et que l'heuristique s'applique bien et est en accord avec la définition réelle de la popularité, nous avons établi les classement des utilisateurs en fonction de chaque champ **_ count*. Les classement ne sont pas identique

TABLE 2: Classement des 10 utilisateurs les plus populaires sur la plateforme Twitch.

Position	Utilisateur	Popularité (entre 0 et 4)
1	eghuk	3.09
2	mineskitv	2.38
3	towelliee	1.40
4	incontroltv	1.31
5	lordkat	1.27
6	chustream	1.25
7	dansgaming	1.00
8	peacefuljay	0.98
9	z60411	0.95
10	redbullgaming	0.85

mais sont assez semblable. En effet, les utilisateurs du top 10 du classement par rapport à notre heuristique se retrouvent également bien classés dans les autres classements.

3.1 Évolution temporelle des flux les plus visionnés

En visualisant l'évolution de l'audience d'un flux au cours du temps, nous avons pu observer que la majorité des flux ont une audience plutôt constante. Nous avons visualisé dans la figure 6, l'évolution temporelle des 20 flux les plus *viraux*, c'est-à-dire ceux qui ont eu le plus grand maximum de spectateurs simultanés.

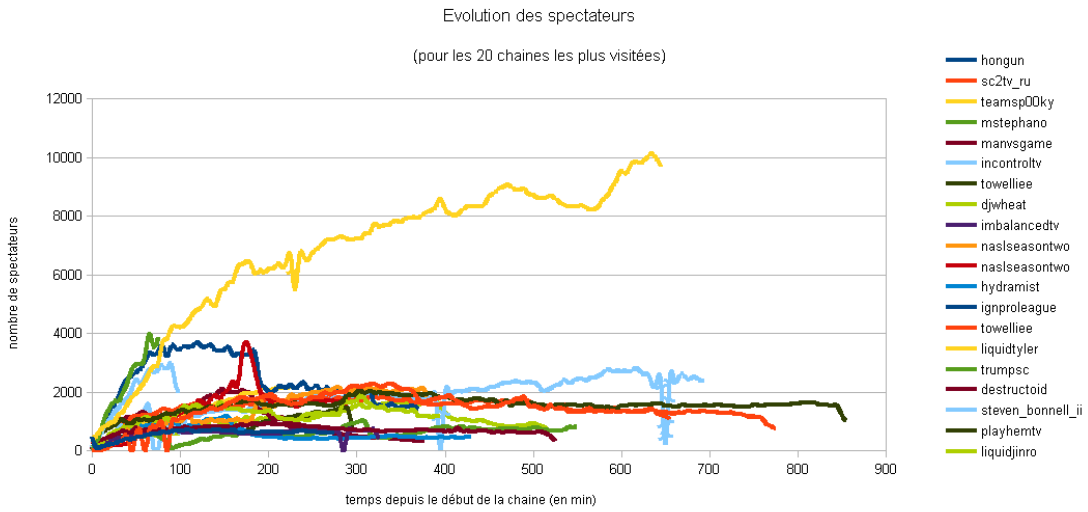

FIGURE 6: Évolution temporelle des flux les plus visionnés

Dans cette dernière, nous avons étudié en détail le plus viral des 20, un flux nommé **Grab Bag of Things**, qui est introuvable par simple recherche. Nous pensons que celui-ci est un véritable *outlier*, sûrement un flux utilisé par les développeurs pour tester la réactivité des serveurs face à une audience accrue.

4 Conclusion

Les difficultés rencontrées ont été variées sur ce projet. Premièrement, comme les données étaient en deux parties bien distinctes sous des formats différents, il a fallu les traiter pour pouvoir tenter de les fusionner. Cela n'a pas été vraiment possible et nous avons donc dû travailler avec deux sources de données différentes. De plus, comme la taille des données est vraiment élevée, il n'a pas été possible de traiter toutes les données. Les résultats annoncés précédemment ne sont donc que des échantillons des résultats totaux.

De plus la plupart des données étaient des classes de valeurs discrètes. Ce n'est donc pas possible de trouver des clusters et donc l'apprentissage et la prédiction d'un critère (la popularité) n'était donc pas possible. Cependant, avec la taille des données, la classification et la visualisation de traits ont été plus aisées.

On a donc pu établir quelques règles et quelques conditions pour pouvoir prédire l'audience d'un flux. Mais cela se fait qu'avec de la lecture et interprétation des données traitées plutôt que par clusterisation.