Un classifieur du comportement des utilisateurs dans les applications pair-à-pair de streaming vidéo

Ihsan Ullah* — **Grégory Bonnet**** — **Guillaume Doyen*** — **Dominique Gaïti***

* *Université de Technologie de Troyes – UMR CNRS 6279 STMR / ERA*
12, rue Marie Curie – 10000 TROYES – France
ullah@utt.fr, doyen@utt.fr, gaiti@utt.fr

** *Université de Caen Basse-Normandie – UMR CNRS 6072 GREYC / MAD*
Boulevard du Maréchal Juin BP 5186 – 14032 Caen Cedex – France
gregory.bonnet@unicaen.fr

RÉSUMÉ. *Depuis quelques années, les applications de streaming vidéo pair-à-pair sont devenues de plus en plus populaires. Cependant, ces systèmes souffrent toujours de problèmes de performance, notamment du fait que, comme les pairs dépendent les un des autres, leur comportement individuel a une influence directe sur le réseau. Pour cette raison, l'étude du comportement des utilisateurs est nécessaire pour définir des mécanismes de contrôle adaptatifs. D'importantes campagnes de mesures ont été réalisées afin d'identifier un comportement moyen conduisant à un modèle général des utilisateurs. Toutefois, le comportement individuel ne suit pas nécessairement le comportement global et les modèles globaux ne sont pas à même de traiter des utilisateurs ayant des habitudes, des préférences et des comportements différents. Nous proposons de tels modèles inspirés de personnages de fictions et nous proposons un classifieur bayésien permettant à un pair de savoir quel modèle lui correspond à partir d'observations.*

ABSTRACT. *P2P-based live video streaming has become popular in recent years. Nevertheless, these systems still suffer from some performance problems such as the fact that, since peers depend upon each other, activities of users have a direct impact over these systems. Therefore, studying the user behavior is needed to design adaptive constrol mechanisms. Several intensive measurement studies have been performed over them in order to provide insights into an average user behavior that can be used for deducing a generalized model. However, the individual user behavior does not necessarily follow the global behavior and a global model is not suitable for dealing with users having different preferences, interests, habits and behaviors. We propose such models of users inspired from fictional characters who represent different kind of consistent behaviors, and we propose a Bayesian classifier that allows a peer to predict in which class it is through simple observations.*

MOTS-CLÉS : *applications multimédia, apprentissage automatique, comportement des utilisateurs, réseaux pair-à-pair.*

KEY WORDS: *machine learning, multimedia applications, peer-to-peer networks, user behavior.*

1. Introduction

Depuis quelques années, les applications de *streaming* vidéo fondées sur une architecture pair-à-pair sont devenues de plus en plus populaires. D'une part et contrairement au *multicast* IP, ces applications ne nécessitent pas de modifications profondes de l'infrastructure du réseau. D'autre part, ces applications réduisent le besoin de déployer de nouveaux serveurs à mesure que le nombre d'utilisateurs du réseau augmente. Une architecture pair-à-pair permet à des hôtes, appelés pairs, de s'auto-organiser au sein d'un réseau virtuel, appelé *overlay*. Les pairs de l'*overlay* partagent leur puissance de calcul et leur bande-passante pour mettre en cache et se diffuser le contenu vidéo les uns aux autres. Toutefois, ces systèmes souffrent toujours de problèmes de performance, notamment en termes de délais à l'initialisation ou lors d'un rembobinage. De plus, comme les pairs dépendent les un des autres, le comportement d'un pair a un effet direct sur le réseau. Par exemple, le départ d'un pair peut entraîner une rupture dans le flux diffusé, diminuant ainsi la qualité de service pour les autres pairs. Pour cette raison, l'étude du comportement des utilisateurs, et par extension des pairs, est nécessaire pour définir des mécanismes de contrôle adaptatifs.

Dans cette optique, d'importantes campagnes de mesures ont été réalisées afin d'identifier un comportement moyen qui peut être utilisé pour déduire un modèle général des utilisateurs. Toutefois, des travaux se heurtent à la difficulté d'établir un lien entre les traces obtenues et des utilisateurs particuliers en raison de l'utilisation d'adresses IP et d'identifiants de pairs dynamiques, de la présence de NAT et de politiques de confidentialité. D'autre part, dans un réseau réel, le comportement individuel ne suit pas nécessairement le comportement global. Ainsi, les modèles globaux ne sont pas à même de traiter des utilisateurs ayant des habitudes, des préférences et des comportements différents. C'est pourquoi, il est nécessaire de développer des modèles individuels pour les intégrer à des outils de simulation ou de gestion de ressources et de qualité de service. Dans cet article, nous proposons de tels modèles inspirés de personnages de fiction qui représentent différents types de comportements individuels cohérents, et nous proposons un classifieur bayésien permettant à un pair de savoir quel modèle lui correspond à partir d'observations.

Cet article est structuré comme suit. Nous présentons dans la Section 2 les travaux relatifs à la modélisation du comportement des utilisateurs, et plus particulièrement l'apprentissage de comportements individuels. De là, nous identifions dans la Section 3 les principaux critères permettant d'identifier les types d'utilisateurs et proposons un réseau bayésien pour classer ceux-ci. Nous identifions en Section 4 plusieurs classes d'utilisateurs et proposons un modèle semi-markovien non homogène pour simuler leur comportement. À partir de ces simulations, nous effectuons l'apprentissage des classes sur notre réseau bayésien et donnons de premiers résultats en Section 5.

2. Travaux relatifs

Depuis quelques années, de nombreuses campagnes de mesures à large échelle ont été rálisée sur des applications de *streaming* vidéo [BRA 99, ACH 04, VIL 05, YU 06, BRA 07, CHA 08b]. Le tableau 1 présente un panorama des modèles globaux qui ont été extraient de ces mesures. Tous s'accordent sur le fait que la population d'un réseau varie selon un cycle journalier [ACH 04, VIL 05, YU 06, CHA 08b] avec un pic en milieu de semaine et [ACH 04] et une décrue durant les week-ends [ACH 04, VIL 05]. Concernant les modèles d'arrivée, les lois exponentielles [ACH 04, CHA 08b]

et les lois de Poisson [BRA 99, YU 06] sont les plus utilisées tandis que, pour les durées de sessions, les modèles se fondent sur des lois lognormales [BRA 99, YU 06, BRA 07] ou exponentielles [VIL 05, CHA 08b]. Concernant les modèles de popularité, toutes les études [ACH 04, VIL 05, YU 06, CHA 08b] s'accordent sur des lois de Zipft bien qu'elles ne modèlisent qu'approximativement les valeurs extrêmes.

Ref.	**Loi d'arrivée**	**Loi de durée de session**	**Loi de popularité**
[BRA 99]	Poisson ($\lambda = 0.68$)	lognormale	non mesurée
[ACH 04]	exponentielle	exponentielle	Zipf ($\alpha = 0.27$)
[VIL 05]	non mesurée	lognormales ($\mu = (0.16, 0.2)$, $\sigma = (0.06, 0.27)$)	Zipf ($\alpha = 0.667$)
[YU 06]	pseudo-Poisson ($\lambda = 17$, $N = 27$)	lognormale ($\mu = 2.2$, $\sigma = 27$)	Zipf
[BRA 07]	non mesurée	lognormale ($\mu = 4.835$, $\sigma = 1.704$)	normale ($\mu = 33.2$, $\sigma = 17.1$)
[CHA 08b]	combinaison d'exponentielles	exponentielle	Zipf

Tableau 1. *Panorama des modèles globaux dans la littérature*

Toutefois, ces travaux s'intéressent uniquement à la définition de modèles globaux d'utilisateurs et non pas à l'identification de leur comportement individuel qui peut être très éloigné de comportement moyen. D'autres travaux se sont alors intéressés à l'identification de critères de comportements et à leur prédiction. [TAN 06] ont mesuré la stabilité des utilisateurs à la fois dans des architectures pair-à-pair mais aussi sur des architectures client-serveur. Ils ont alors identifié une corrélation positive entre le temps passé dans le syst`eme et la durée de session restante d'un utilisateur. Ils proposent un modèle de sélection de voisins tel que les pairs présents depuis le plus de temps dans le réseau sont préférés aux autres. [WAN 08] propose une méthode similaire pour identifier des pairs stables. Afin de minimiser l'effect de l'attrition, ils se servent ensuite de ces pairs comme ossature du réseau. Toutefois, ces deux approches ont tendance à considérer comme instables tous les pairs récemment arrivés dans le réseau, ce qui diminue la qualité de service de ces derniers, et favorise donc leur instabilité. [LIU 09a] ont alors mesuré et analysé l'effet de la qualité du flux sur la stabilité et la contribution en bande-passante des pairs. Ils observent que cette qualité présente une corrélation positive avec ces deux critères. Ils proposent alors un modèle qui reste toutefois un modèle global et ne considère que des utilisateurs homogènes.

Afin de définir des modèles individuels, [HOR 09] proposent plutôt d'apprendre la contribution en bande-passante des pairs à l'aide d'une machine à vecteur support (SMV) ; mais ils ne considèrent aucune autre crtière de comportement. Dans nos travaux précédents [ULL 09], nous avons proposé des estimateurs de la stabilité des pairs fondés sur une moyenne mobile exponentielle et la règle de Bayes. De plus, nous avons proposé un mécanisme proactif pour anticiper le départ de voisins et en sélectionner de nouveaux dynamiquement. Toutefois, ce travail ne se fonde que sur l'historique des

pairs. C'est pourquoi, dans [ULL 10], nous avons proposé une modélisation fondée sur un réseau bayśien pour tenir compte de l'influence de tous les critères de comportements. Ce travail présente toutefois deux limites : il est difficile d'obtenir des estimations avec une granularité satisfaisante, et l'apprentissage du modèle de l'utilisateur nécessite de très nombreuses observations.

Afin de pallier ces limites, nous proposons de non plus apprendre le modèle individuel des utilisateurs mais, sachant l'existence préalable d'un ensemble de modèles types, d'étiquetter chaque utilisateur par le modèle qui approxime au mieux son comportement.

3. Classification des utilisateurs

Dans un premier temps, nous identifions à partir de la littérature les métriques qui influencent la classification des utilisateurs. De là, nous proposons un classifieur fondé sur un réseau bayésien.

3.1. *Identification des critères*

Le principal critère d'identification d'un utilisateur est le temps qu'il passe sur un canal donné sous certaines condition. Ce critère, appelé durée de session, est en effet la seule observation qu'un pair peut avoir d'un autre, et par conséquent la seule observation sur laquelle il est possible de se fonder pour identifier un pair. Dans la littérature, quatre métriques influencent la durée de session :

1) **la qualité du flux :** [LIU 09a, LIU 09b] ont identifié une corrélation positive entre la durée de session et la qualité initiale du flux. Cette qualité de flux est évaluée par la taille du tampon de donnée reųt initialement par l'utilisateur. En effet, un utilisateur qui reçoit un important tampon initial a tendance à rester plus longtemps connecté au réseau ;

2) **la popularité :** [LIU 09a, HEI 07, LIU 09b] observent que les utilisateurs ont tendance à présenter une durée de session plus longue lorsqu'ils regardent un programme populaire, et inversement pour les programmes non populaires. Dans les applications de diffusion de contenu, la popularité est calculée en fonction de la population présente dans le réseau.

3) **le type de programme :** [CHA 08a] observent des durées plus courtes pour les programmes d'information et de musique comparativement aux documentaires et programmes pour les enfants. Ces auteurs ont ainsi pu identifier trois types de contenu qui entraînent des comportements différents : fiction (films, séries, programmes pour enfant), réalité (informations, documents) et sports ;

4) **l'heure de la journée :** [LIU 09a] montrent que la durée de session est fortement corrélée à l'heure de la journée.

Ainsi, nous pouvons classifier le comportement d'un utilisateur en fonction de sa durée de session sachant les valeurs de ces quatres métriques. Pour cela, nous proposons d'utiliser un réseau bayésien.

3.2. *Un classifieur bayésien*

Le classifieur que nous proposons est représenté par le réseau bayésien sur la figure 1. Il possède six nœuds : un pour la durée de session, quatre pour les métriques précédemment identifiées et un

représentant la classe de l'utilisateur. La durée de session présente un arc vers la classe d'utilisateur, et les quatre métriques présentent un arc vers la durée de session et vers la classe de l'utilisateur.

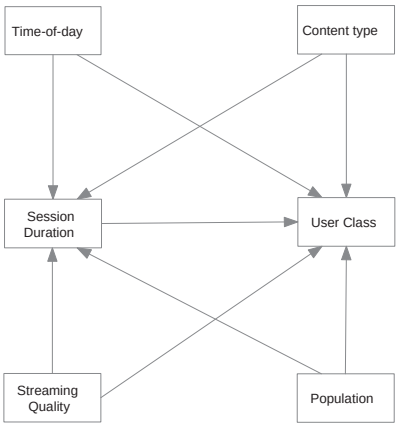

Figure 1. *Diagramme du classifieur bayésien*

Nous considérons chaque variable représentée par un nœud comme discrète. L'heure de la journée est discrétisée en 24 états, un pour chaque heure. Le type de contenu est discrétisé en 3 états : fiction, réalité et sports. La qualité du flux et la popularité sont toute deux discrétisées en cinq états selon l'équation 1 : S est la valeur de létat calculé, V_{courant} est la popularité ou la qualité actuelle du canal tandis que V_{max} en est la valeur maximale.

$$S = \left\lceil 5. \frac{V_{\text{courant}}}{V_{\text{max}}} \right\rceil \tag{1}$$

La durée de session est, elle, discrétisée en 100 états d'une granularité d'une minute afin de couvrir toute les durées de session possibles. En dernier lieu, la classe de l'utilisateur est un nœuds comprenant autant d'état que de catégories d'utilisateurs identifiées dans l'application cible. Dans notre cas, six classes d'utilisateurs ont été définies et sont présentées dans la section suivante.

4. Définition des classes d'utilisateurs

Malgré l'existance de modèles globaux d'utilisateurs, il est difficile d'en dériver des modèles individuels. Pour cette raison, il est intéressant de faire appel à une technique venant de l'étude de marché : la définition de personas. Cela consiste à définir des personnages de fiction cohérents en termes d'identité, de compétences, d'habitudes et de buts [GRU 06]. Par exemple, dans le cadre du projet On-Demand[1] qui a pour objectif d'améliorer la qualité de service sur les réseaux de diffusion de contenu, [RUD 08] utilisent cette méthodologie pour définir des archétypes de téléspectateurs.

1. http ://redback.sics.se/projects/ondemand_iptv

Toutefois, ce travail ne présente qu'une description qualitative comme indiqué dans la première partie de la table 2. Nous avons alors utilisé ces personas comme classes d'utilisateurs et nous en proposons un modèle quantitatif pouvant être simulé.

4.0.1. *Un modèle semi-markovien non homogène*

Chaque utilisateur est modélisé par un processus semi-markovien non homogène. Cela signifie que l'état d'un utilisateur (en ligne ou hors ligne) à un instant donné dépend non seulement de son état à l'instant précédent comme pour tout modèle markovien mais aussi du temps qu'il a passé dans l'état courant et du temps global du processus. Cette catégorie de processus s'adapte bien au contexte du *streaming* vidéo car la littérature montre que le comportement d'un utilisateur varie en fonction de l'heure de la journée et du temps qu'il a déjà passé dans le réseau [LIU 09a].

Nous faisons les hypothèses suivantes :

1) Nous considérons des applications mono-canal et définissons donc deux états $\{X_1, X_2\}$ dont la sémantique correspond respectivement à la présence en ligne et hors-ligne de l'utilisateur ;

2) Nous considérons un comportement cyclique sur une journée et définissons le temps global du processus $t \in \mathbb{N}^+$ discrétisé en intervalles d'une minute, $\{t_1 \dots t_{1440}\}$;

3) Afin d'être cohérent avec [BRA 99, VIL 05, YU 06, BRA 07] et en raison du fait que les lois lognormales sont couramment utilisées pour modéliser des processus qui s'évanouissent lentement, nous définissons la transition de X_1 en X_2 comme contrôlée par une loi log-normale. Ainsi, les paramètres $\mu_t^i(d)$ et $\sigma_t^i(d)$ de cette loi dépendent du persona i, du type de contenu d, du temps global t et de t_{X_1} le temps passé en X_1 ;

4) Comme les lois de Poisson sont couramment utilisée pour modéliser des processus d'arrivée, nous définissons la transition de X_2 en X_1 comme contrôlée par une telle loi, ce qui est cohérent avec [BRA 99, YU 06]. Ainsi, le paramètre λ_t^i de cette lo dépend du persona i et du temps glboal du processus t.

L'équation 2 donne la probabilité de transition de X_1 vers X_2. Il s'agit de la probabilité d'être déconnecté après un temps t_{X_1} passé en X_1.

$$P(X(t) = X_2 | X_1, t_{X_1}) = \int_0^{t_{X_1}} \frac{e^{\frac{-(\log x - \mu_t^i(d))^2}{2.(\sigma_t^i(d))^2}}}{x.\sigma_t^i(d).\sqrt{2\pi}}.dx \tag{2}$$

L'équation 3 donne la probabilité de transition de X_2 vers X_1. Il s'agit de la probabilité d'être connecté à un moment donné de la journée. Comme nous nous intérrons à un unique utilisateur, nous fixons le $k = 1$ pour la loi de Poisson.

$$P(X(t) = X_1 | X_2) = e^{-\lambda_t^i} \lambda_t^i \tag{3}$$

Dans toute la suite, nous notons les personas de [RUD 08] par $\{J, E, S, A, P, L\}$, c'est-à-dire **J**ohnatan, **E**mma, **S**tephan, **A**nna, **P**eter et E**L**len. Nous les instancions ensuite les valeurs des paramètres comme indiqué dans la table 2.

4.0.2. *Modèle d'arrivée*

En accord avec les observations de [ACH 04, VIL 05, YU 06, CHA 08b], nous définissons cinq périodes de la journée : matin, midi, après-midi, soirée et nuit. Pour chaque persona et chaque période, nous définissons le λ_t^i sur un intervalle d'une minute (cf. la deuxième partie de la table 2). Nous choisissons ces valeurs de manière à refléter les habitudes de chaque persona. Par exemple, J est présent dans le réseau plus souvent la soirée que le matin. De plus, la somme de ces valeurs est inférieure à 1, signifiant qu'un utilisateur donné ne se connecte pas nécessairement au réseau tous les jours. Enfin, la somme de tous les λ_t^i permet d'approximer l'évolution globale de la population.

4.0.3. *Modèle de durée de session*

La première partie de la table 2 nous indique que chaque persona est présent dans le réseau durant un temps moyen chaque jour. Selon [RUD 08], J et A sont très réguliers, E et L sont réguliers, P est irrégulier et S est très irrégulier. Cette régularité correspond à la variance $\sigma^2 = \frac{\bar{x}}{k}$ où $\bar{x}$ est le temps de présence par jour ($k = 1$ pour très irrégulier à 4 pour très régulier). La troisième partie de la table 2 donne la moyenne et la variance en minutes de ce temps pour chaque persona i ainsi que les μ_i et σ_i des lois correspondantes.

Comme [RUD 08] ne fournit que le temps de présence par jour, nous devons définir $\mu_t^i(d)$ et $\sigma_t^i(d)$ à partir de μ_i et σ_i. Nous faisons l'hypothèse que la durée de session au cours d'une période donnée de la journée correspond à une portion du temps de présence total multipliée par un cœfficient d'intêret en fonction du type de contenu. La quatrième partie de la table 2 indique cette portion pour chaque persona en fonction de la période. Les valeurs de ces paramètres sont arbitraires mais reflètent les habitudes de chaque persona. La cinquième partie de la table 2 indique le cœfficient d'intêret θ_d^i pour chaque persona i en fonction du contenu d. Nous avons choisi les valeurs des cœfficient telles que leur moyenne égale 1, ce qui représente le fait qu'en moyenne la durée de session espérée suit la distribution initiale. Par conséquent, l'équation 4 définit $\mu_t^i(d)$ et $\sigma_t^i(d)$.

$$\begin{aligned} \mu_t^i(d) &= \mu_i.M_i^t.\theta_d^i \\ \sigma_t^i(d) &= \sigma_i.M_i^t.\theta_d^i \end{aligned} \tag{4}$$

4.0.4. *Modèle de population*

En tant que modèle individuel, un processus semi-markovien non-homogène est instancié pour chaque utilisateur du réseau. Toutefois, certaines classes de comportements sont plus courantes que d'autres car il y a moins par exemple moins de retraîtées (L) que de jeunes hommes (J) au sein d'une population quelconque. Ainsi, nous proposons de fixer la proportion des persona en fonction de donnée obtenues auprès de l'Institut National de la Statistique et des Études Économiques (INSEE). La sixième partie de la table 2 indique cette répartition.

5. Évaluation

Afin de valider notre proposition, nous avons réalisé des simulations à l'aide de Matlab et de la Bayes Net Toolbox.

Paramètre \ Persona	**Johnatan** (J)	**Emma** (E)	**Stephan** (S)
Caractéristiques définies par [RUD 08]			
Âge	17	25	33
Intérets	Sports et séries	Sans préférences	Nouvelles et sports
Temps de présence par jour	2 - 3 h.	1.5 h.	1.5 h. (Δ important)
Habitudes	Soirée et nuit	Nuit	Midi et soirée
Catégorie	Étudiant	Commerçant	Cadre
Valeur de λ_i^t sur un intervalle d'une minute en fonction du persona i et du temps global t			
4 :00 à 9 :59	0.00125	0.00125	0.00125
10 :00 à 15 :59	0.00375	0.00375	0.0075
16 :00 à 18 :59	0.00375	0.00125	0.00125
19 :00 à 22 :59	0.0075	0.005	0.00375
23 :00 à 3 :59	0.00375	0.00625	0.00375
Valeurs de μ_i et σ_i sur une journée complète en fonction du persona i			
Temps passé en ligne par jour	150	90	90
Variance	37.5	30	90
μ_i	5.0098	4.498	4.4943
σ_i	0.0408	0.0608	0.1051
Proportion M_t^i du temps de présence en fonction du persona i et du temps global t			
4 :00 à 9 :59	0.05	0.05	0.05
10 :00 à 15 :59	0.1	0.05	0.2
16 :00 à 18 :59	0.1	0.05	0.1
19 :00 à 22 :59	0.4	0.2	0.4
23 :00 à 3 :59	0.35	0.65	0.25
Valeur des cœfficient d'intêret θ_d^i en fonction du persona i et du type de contenu d			
Fiction	1	1	0.3
Réalité	0.5	1	1.7
Sports	1.5	1	1
Proportion de chaque persona dans le réseau			
	0.182	0.168	0.177

Paramètre \ Persona	**Anna** (A)	**Peter** (P)	**Ellen** (L)
Caractéristiques définies par [RUD 08]			
Âge	46	58	69
Intérets	Séries	Nouvelles et reportages	Variété et reportages
Temps de présence par jour	1.5 h.	1.5 h.	2 h.
Habitudes	Après-midi et soirée	Midi et soirée	Sans habitudes
Catégorie	Femme au foyer	Professeur	Retraitée
Valeur de λ_i^t sur un intervalle d'une minute en fonction du persona i et du temps global t			
4 :00 à 9 :59	0.00375	0.00125	0.00375
10 :00 à 15 :59	0.005	0.005	0.005
16 :00 à 18 :59	0.00625	0.00375	0.00375
19 :00 à 22 :59	0.00625	0.0075	0.00375
23 :00 à 3 :59	0.00125	0.00125	0.00125
Valeurs de μ_i et σ_i sur une journée complète en fonction du persona i			
Temps passé en ligne par jour	90	90	120
Variance	22.5	45	40
μ_i	4.4984	4.497	4.7861
σ_i	0.0527	0.0744	0.0527
Proportion M_t^i du temps de présence en fonction du persona i et du temps global t			
4 :00 à 9 :59	0.05	0.05	0.25
10 :00 à 15 :59	0.2	0.3	0.25
16 :00 à 18 :59	0.35	0.2	0.15
19 :00 à 22 :59	0.2	0.35	0.25
23 :00 à 3 :59	0.2	0.1	0.1
Valeur des cœfficient d'intêret θ_d^i en fonction du persona i et du type de contenu d			
Fiction	2.4	0.3	1
Réalité	0.3	2.4	1.5
Sports	0.3	0.3	0.5
Proportion de chaque persona dans le réseau			
	0.186	0.174	0.113

Tableau 2. *Paramètres du modèle*

Paramètre	Valeur
Durée de la simulation	60 jours (apprentissage) 40 jours (validation)
Population	1000 pairs
Contenu	fiction, réalité, sports
Durée d'un programme	2 heures
Algorithme d'inférence	Arbre de jonction
Algorithme d'apprentissage	Maximum de vraisemblance

Tableau 3. *Paramètres de la simulation*

5.1. *Paramètres des simulations*

Les paramètres de nos simulations sont résumés dans la table 3. Nous avons ainsi généré des traces pour 1000 utilisateurs distincts qui ont utilisé un réseau sur une période de 100 jours. Les 60 premiers jours ont servi à fournir des exemples pour l'apprentissage tandis que les 40 autres jours ont été dévolus à l'évaluation du classifieur.

5.2. *Premiers résultats*

Nous utilisons le classifieur bayésien pour prédire la classe d'utilisateur d'un pair donné. Le réseau bayésien met à jour ses tables de probabilités conditionnelles après chaque observation du comportement d'un pair. Durant les 60 premier jour, le classifieur réalise un apprentissage supervisé non bruité : toutes les variables sont observables. À partir du 61^{e} jour, la variable de classe d'utilisateur est cachée et le classifieur prédit cette dernière comme étant celle qui maximise la probabilité d'être sachant les valeurs des autres variables.

La matrice de confusion obtenue est donnée dans la table 4. Afin de clarifier les résultats, la table 5 indique l'erreur obtenue. Nous pouvons remarquer que pour la moitiée des classes d'utilisateurs, l'erreur reste en-dessous de 10%. Cependant, elle atteint $23,5\%$ pour les autres. La figure 2 trace la droite de regression entre les classifications correctes et le nombre total de classifications.

Ces résultats sont bons pour certaines classes mais pas pour toutes. La principale raison de cette erreur vient du fait que le classifieur doit nécessairement classer un pair dans une et unique classe, quelle que soit l'information qu'il possède. Il est des cas où la distribution de probabilité entre les classes est par trop homogène pour donner une prédiction pertinente.

5.3. *Amélioration de la précision*

Pour les applications où la précision prime sur le fait de classer tous les utilisateurs, décider entre deux classes nést pas nécessairement pertinent. Par exemple, nous pouvons considérer un protocole de construction d'une ossature de réseau à partir d'utilisateur stable. Pour un tel protocole, il est important d'identifier correctement de tels utilisateurs mais il n'est pas nécessaire de construire

Classe réelle	**Classe prédite**					
	J	E	S	A	P	L
J	30128	645	381	619	359	758
E	957	24488	714	205	198	221
S	767	3277	26686	459	2146	1563
A	743	906	558	39030	303	957
P	476	631	1695	1375	26174	577
L	436	219	730	1920	532	16980

Tableau 4. *Matrice de confusion*

Classe	**# cas**	**# erreur**	**% erreur**
J	32890	2762	$8,3977$
E	26783	2295	$8,5689$
S	34898	8212	$23,5314$
A	42497	3467	$8,1582$
P	30928	4754	$15,3712$
L	20817	3837	$18,4321$

Tableau 5. *Erreur de classification*

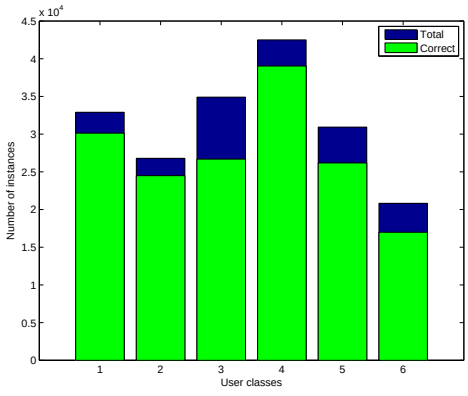

Figure 2. *Classifications selon les différents personas*

l'ossature à partir de tous les utilisateurs stables du système : il suffit d'identifier un simple sous-ensemble.

Seuil	Erreur pour chaque classe						Inconnue (%)
	J	E	S	A	P	L	
$0,2$	$8,4424$	$8,1918$	$23,0397$	$7,8177$	$14,9615$	$17,8321$	$0,5111$
$0,3$	$8,2823$	$7,9772$	$22,8758$	$7,6250$	$14,7406$	$17,7039$	$1,3145$
$0,4$	$6,8627$	$7,5310$	$21,8501$	$7,3008$	$14,1242$	$16,5759$	$3,5474$
$0,5$	$5,8305$	$6,3165$	$18,8363$	$6,4280$	$11,7597$	$13,8887$	$8,4406$
$0,6$	$4,0089$	$5,8304$	$8,0827$	$4,4150$	$9,2384$	$12,4707$	$17,8081$
$0,7$	$2,9120$	$4,2268$	$6,4121$	$3,9540$	$7,9906$	$12,0455$	$22,7278$
$0,8$	$1,7908$	$3,6179$	$4,4427$	$2,7311$	$2,4767$	$7,4937$	$34,8954$
$0,9$	$0,7829$	$2,3012$	$2,4788$	$0,7595$	$0,8710$	$2,3440$	$51,9683$

Tableau 6. *Évolution de l'erreur sous différents seuils*

Nous proposons d'améliorer le classifieur pour de telles applications en introduisant une nouvelle classe d'utilisateur, appelée *inconnue*, et en fixant un seuil de précision dans l'intervalle $[0, 1]$. Si la probabilité d'appartenance à la classe prédite est en-dessous de ce seuil alors le classifieur indique que la classe est inconnue. Nous avons essayé plusieurs niveaux de seuil, et indiquons en table 6 l'erreur pour chaque classe et le nombre d'inconnus. Nous pouvons alors remarquer que l'erreur pour chaque classe décroît au fur et à mesure que le seuil augmente, au prix d'une augmentation du nombre d'inconnus. Nous pouvons nous servir de cette approche pour fixer un compromis entre précision et classification totale.

6. Conclusions

Modéliser de manière précise les utilisateurs d'une application pair-à-pair de *streaming* vidéo ouvre de nombreuses perspectives, depuis son intégration à des outils de simulations à la gestion de ressources pour les fournisseurs de services en passant par la mise en œuvre de systèmes permettant d'améliorer la qualité de service. Toutefois, les modèles globaux proposés dans la littérature s'éloignent du comportement individuel des utilisateurs. Dans cet article, nous proposons un classifieur bayésien permettant d'associer à chaque utilisateur une classe de comportement. Afin de valider ce classifieur, nous définissons à l'aide de la méthodologie des personas six classes d'utilisateurs. Nos premier résultats sont mitigés : la moitié des classes est prédite avec une précision satisfaisante. Nous proposons alors l'introduction d'un biais permettant de classer avec une plus grande précision les utilisateurs au détriment du nombre de prédiction.

Ce travail doit toutefois être étendu afin d'une part d'améliorer la définition des classes d'utilisateurs, et d'autre part d'intégrer cet outil dans deux applications. La première concerne la construction de cœur de réseau autour d'utilisateurs stables. La seconde concerne l'adaptation automatique de la topologie du réseau en fonction de la proportion de persona présente à un instant donné.

7. Bibliographie

[ACH 04] ACHARYA S., SMITH B., « Characterizing user access to videos on the world wide web », *Lecture Notes in Computer Science*, vol. 2720, 2004, p. 375–384.

[BRA 99] BRANCH P., EGAN G., TONKIN B., « Modeling interactive behaviour of a video based multimedia system », *Proceedings of the IEEE International Conference on Communications*, 1999, p. 978-982.

[BRA 07] BRAMPTON A., MACQUIRE A., RAI I., NICHOLAS J.-P., MATHY L., FRY M., « Characterising user interactivity for sports video-on-demand », *Proceedings of the 17th NOSSDAV*, 2007.

[CHA 08a] CHA M., RODRIGUEZ P., CROWCROFT J., MOON S., AMATRIAIN X., « Watching television over an IP network », *Proceedings of the 8th IMC*, 2008, p. 71–84.

[CHA 08b] CHANG B., DAI L., CUI Y., XUE Y., « On feasibility of P2P on-demand streaming via empirical VoD user behavior analysis », *Proceedings of the 28th ICDCS*, 2008, p. 7–11.

[GRU 06] GRUDIN J., « Why personas work : the psychological evidence », *The Persona Lifecycle : Keeping People in Mind Throughout Product Design*, p. 642–664, Elsevier Inc., 2006.

[HEI 07] HEI X., LIANG C., LIANG J., LIU Y., ROSS K. W., « A measurement study of a large-scale P2P IPTV system », *IEEE Transactions on Multimedia*, vol. 9, n° 8, 2007, p. 1672–1687.

[HOR 09] HOROVITZ S., DOLEV D., « Collabrium : active traffic pattern prediction for boosting P2P collaboration », *Proceedings of the 18th WETICE*, 2009, p. 116–121.

[LIU 09a] LIU Z., WU C., LI B., ZHAO S., « Distilling superior peers in large-scale P2P streaming systems », *Proceedings of the 28th INFOCOM*, 2009, p. 82–90.

[LIU 09b] LIU Z., WU C., LI B., ZHAO S., « Why are peers less stable in unpopular P2P streaming channels ? », *Networking*, 2009, p. 274–286.

[RUD 08] RUDSTROM A., SJOLINDER M., « Capturing TV user behaviour in fictional character descriptions », rapport, October 2008, Swedish Institute of Computer Science.

[TAN 06] TANG Y., SUN L., LUO J.-G., ZHONG Y., « Characterizing user behavior to improve quality of streaming service over P2P networks », *Advances in Multimedia Information Processing*, 2006, p. 175–184.

[ULL 09] ULLAH I., BONNET G., DOYEN G., GAÏTI D., « Improving performance of ALM systems with Bayesian estimation of peers dynamics », *Proceedings of the 12th MMNS*, 2009, p. 157–169.

[ULL 10] ULLAH I., BONNET G., DOYEN G., GAÏTI D., « Modeling user behavior in P2P live video streaming systems through a Bayesian network », *Proceedings of the 4th AIMS*, 2010, p. 2–13.

[VIL 05] VILAS M., PANEDA X.-G., GARCIA R., MELENDI D., GARCIA V.-G., « User behavior analysis of a video-on-demand service with a wide variety of subjects and lengths », *Proceedings of the 31st EUROMICRO Conference on Software Engineering and Advanced Applications*, 2005, p. 330-337.

[WAN 08] WANG F., LIU J., XIONG Y., « Stable peers : existence, importance and application in Peer-to-Peer live video streaming », *Proceedings of the 27th INFOCOM*, 2008, p. 1364–1372.

[YU 06] YU H., ZHENG D., ZHAO B. Y., ZHENG W., « Understanding user behavior in large-scale video-on-demand systems », *Operating Systems Review*, vol. 40, n° 4, 2006, p. 333–344.