RAPPORT DE STAGE DE RECHERCHE

Temporalité dans la recommandation de spectacles

Avril - Août 2018

Cyril Malbranke

INTRODUCTION

Au cours de ma 3e année à l'École, j'ai suivi le Programme d'Approfondissement "Image, Vision et Apprentissage", ce PA m'a donné l'opportunité de m'intéresser en particulier aux différentes techniques de Machine Learning. Au détour de certains cours (NLP en particulier), j'ai été intéressé par les possibilités offertes par l'étude des graphes et particulièrement les graphes relationnels ou sociaux.

C'est pour cette raison que j'ai choisi d'effectuer mon stage au LIP6 dans l'équipe "Complex Networks" qui étudie les propriétés et les capacités de ces réseaux entre autres. Mon stage a été effectué en partenariat avec la Start-Up Delight Data Enternainment, c'est également cette approche plus industrielle qui m'a séduite.

Mon stage portait sur l'étude de la temporalité dans la recommandation de spectacles vivants (théâtre, cinéma, concerts ...). La recommandation dans ce domaine présente de nombreuses difficultés : les volumes de ventes peuvent être assez faibles, et les consommateurs de spectacles ont souvent des profils particulier. Des approches classiques tels le filtrage collaboratif sont donc souvent inapplicables en l'état.

L'objectif est de comprendre comment utiliser au mieux l'information temporelle contenue dans des données de billetteries. En nous basant sur un formalisme innovant qui sera détaillé dans notre rapport, nous chercherons à développer des méthodes hybrides qui tirent parti à la fois de la structure relationnelle entre spectacles et spectateurs et de la série temporelle des achats.

Notre rapport décrira dans un premier temps les outils utilisés dans notre moteur de recommandation et motivera les décisions prises dans la conception de notre système. Dans un deuxième temps nous dresserons une explication détaillée du moteur de recommandation.

CONTENTS

1 PRÉSENTATION

1.1 Présentation des organismes impliqués

• Le Laboratoire d'Informatique de Paris VI

Le **LIP6**, Unité Mixte de Recherche de Sorbonne Université et du Centre National de la Recherche Scientifique, est un laboratoire de recherche en informatique se consacrant à la modélisation et la résolution de problèmes fondamentaux motivés par les applications, ainsi qu'à la mise en œuvre et la validation des solutions au travers de partenariats académiques et industriels. Il emploie plusieurs centaines de personnes et couvre un large spectre de domaines allant de la science des données à l'étude des réseaux et systèmes.

L'équipe dans laquelle j'ai effectué mon stage est l'équipe **Complex Network**, elle étudie les propriétés de graphes "réels", connectés à de nombreux problèmes théoriques et appliqués. Ces graphes ont des propriétés communes et sont donc intéressant à étudier dans leur globalité. L'équipe aborde différents problèmes : extraction de données de ces graphes, description de leur structure dynamique et global, algorithmes pour traiter les grands graphes ... Les problèmes abordées ouvrent de plus la voie à une coopération inter-disciplinaire avec d'autres groupes du laboratoire ou d'autres organismes.

• la start-up Delight Data Entertainment

Delight est une plateforme qui propose des outils de marketing variés aux producteurs et aux salles de spectacle. Son objectif est de les aider à mieux comprendre les motivations des gens pour mieux cibler les actions de communication. Elle vise ainsi à mieux appréhender le segment-client très particulier du monde du spectacle.

Delight a déjà développé une plateforme de traitement, nettoyage et visualisation de données qui lui a permis d'obtenir la confiance et le partenariat de nombreux (grands) partenaires dans le monde du spectacle. Elle a ainsi pu récolter des volumes de données exploitables suffisants pour pouvoir envisager le développement d'outils de traitements de données plus complexes. Parmi ses projets on trouve notamment la conception d'algorithmes de recommandations et de mise en relation clients-spectacles : tant pour la recommandation de spectacles susceptibles d'intéresser un client, que pour la recommandation à des producteurs de clients ou de segments de clients à "cibler".

1.2 Présentation et objectifs du projet

• Présentation du projet

Mon stage s'inscrit dans la continuité du développement de la théorie sur les flots de liens au sein du LIP6. Les points essentiels seront développés plus tard dans notre rapport, mais le lecteur est invité à se référer à la bibliographie pour de plus amples développements [5].

Les **flots de liens** sont un formalisme mathématique développé afin de faciliter l'exploitation combinée des propriétés structurelles et dynamiques des graphes. Ils sont donc un outil très naturel pour la prédiction de formation de liens, dont les applications sont multiples. Sa formalisation sera développé plus en avant de notre rapport (section 2.2).

Dans l'objectif d'étudier le potentiel de cet outil mathématique dans le domaine de la prédiction, un travail de thèse a été réalisé par Thibaud Arnoux [1], dans l'objectif de combiner les données structurelles et dynamiques d'un graphe pour prédire l'activité future d'un système modélisé sous cette forme. Son étude se concentrait sur des jeux de données d'interactions entre personnes. Ces jeux de données sont la plupart du temps construits au cours de conférences ou dans des universités, écoles ou lycées en équipant les personnes présentes de détecteurs qui enregistrent sur une période donnée les interactions entre les personnes. L'objectif à partir de ces données est donc de prédire les liens qui se formeront sur une période ultérieure. Ses travaux seront développés et contextualisés dans la section 2.3

• Objectifs du stage

L'objectif de notre stage est donc de s'appuyer sur ces travaux de thèses. Notre objectif est d'étudier dans quelle mesure la combinaison des informations dynamiques et structurelles telles que proposer par les flots de liens peut être pertinente dans un objectif de recommandation.

Dans la suite de nos travaux, nous formaliserons de la manière suivante le problème de la recommandation : soit un graphe bipartie, avec d'un côté des nœuds *clients* et de l'autre des nœuds *spectacles*. A partir des interactions sur une période définie entre ces nœuds (Interaction = un client achète une place pour un spectacle), nous souhaitons prédire les interactions à venir dans le graphe clients-spectacles.

Pour ce faire, nous réaliserons une adaptation des travaux de Thibaud Arnoux à notre cas d'étude. Les problèmes soulevées par ce cas sont nombreux :

- **Identification de comportements temporels** utiles à un travail d'assimilation et de différenciation sur les spectacles et sur les clients. Cette problématique fera office de travail préliminaire et sera développée dans la partie 3

- **Établissement des nœuds *spectacles* et *clients***. Il n'y a *a priori* pas de récurrence de liens entre un client individuel et un spectacle individuel. Il paraît donc nécessaire de créer cette récurrence de liens. Cette problématique sera abordée dans les sections 4.1 et 4.2
- **Adaptation des travaux** [1] au problème particulier de la recommandation, cette problématique sera abordée dans la section 4.3.
- **Utilisation de modèles, fonctions et techniques d'apprentissage et de prédictions plus complexes**, pour espérer améliorer nos résultats. Cet axe sera travaillé dans la suite de mon stage et ses embryons sont discutés partie 5.

2
ÉTAT DE L'ART

Dans cette partie, nous poserons les bases nécessaires pour la suite de notre travail par une reprise des travaux sur lesquels nous nous appuierons dans les parties suivantes. Dans un second temps nous dresserons une analyse ciblée de la composition de notre jeu de données de travail.

2.1 Filtrage collaboratif et recommandation basé sur le contenu

Les systèmes de recommandation actuelles sont aujourd'hui parcourus par deux conceptions assez dominantes, en se référant à Ricci et al, 2015 [7] :

- La filtrage collaboratif. Schématiquement, on établit la ressemblance entre deux individus, si l'un consomme un objet, on aura tendance à proposer au second ce même objet, si ces individus sont assez similaires.

- La recommandation **basé sur le contenu** (*content based*). On établit la ressemblance entre deux objets sur ses données propres (informations sémantiques par exemple), si l'utilisateur consomme l'un des objets on aura tendance à lui recommander de consommer le second similaire [8].

Au cours de notre stage, nous chercherons à combiner les possibilités de ces deux conceptions. Dans la suite de cette section, nous traitons l'aspect *content based* et comment il sera retranscrit dans la suite de notre travail.

• Tags d'un item et occurrences

Soit I un ensemble d'items. On associe à chaque item $i \in I$, un ensemble de mots caractérisant un item qu'on nomme **tags** (typiquement un ensemble de mots clefs). Notons W l'ensemble des mots utilisés pour la caractérisation d'un item et considérons :

$$\begin{array}{rccc} T & : & I & \to & \mathscr{P}(W) \\ & & i & \mapsto & T(i) \end{array}$$

On définit alors la fonction d'**occurrences** qui à chaque tag associe l'**ensemble** des items dans lesquels il occurre, pour $w \in W$ on note $occ(w) = \{i \in I, w \in T(i)\}$.

• Mesures de similarités basées sur des tags

Plusieurs mesures de similarités peuvent découler de cette notion d'occurrence et seront utilisées dans la suite de nos travaux. Elles serviront aux regroupements d'évènements dans la section 4.1.

- La **cooccurrence** : $coo(v,w) = |occ(v) \bigcap occ(w)|$.
  Le nombre de fois où les tags sont présents tous les deux dans le même évènement.

- Le coefficient de **Jaccard** : $J(v,w) = \frac{|occ(v) \bigcap occ(w)|}{|occ(v) \bigcup occ(w)|}$.
  C'est une normalisation de la cooccurrence. On évite ainsi une sur-évaluation de la similarité de deux tags, si l'un des deux ou les deux sont trop courants

- Le coefficient de **Jaccard asymétrique** : $J_a(v,w) = \frac{|occ(v) \bigcap occ(w)|}{|occ(v)|}$.
  C'est une normalisation asymétrique de la cooccurrence. Elle permet de mieux rendre compte des différences de "hiérarchie". Un exemple intuitif serait un tag *musique*, les tags *rock* ou *rap* seraient très souvent associés à *musique*, alors que le tag *musique* serait relativement beaucoup moins associé à *rock* ou *rap*.

- La mesure **BP** [3] : $BP(v,w) = \frac{coo(v,w) - |occ(v)||occ(w)|/|I|}{\sqrt{|occ(v)|(1-|occ(v)|)|occ(w)|(1-|occ(w)|)}}$.
  Elle quantifie un écart à une hypothèse zéro d'indépendance entre le tag v et le tag w. En effet, sous une hypothèse d'indépendance on a $\forall i$ $\mathbf{P}(v \in T(i) \cap w \in T(i)) = \frac{|occ(v)|}{|I|} \cdot \frac{|occ(w)|}{|I|}$ d'où $\mathbf{E}[coo(v,w)] = \frac{|occ(v)| \cdot |occ(w)|}{|I|}$

2.2 Flots de liens : Définitions et Formalisations

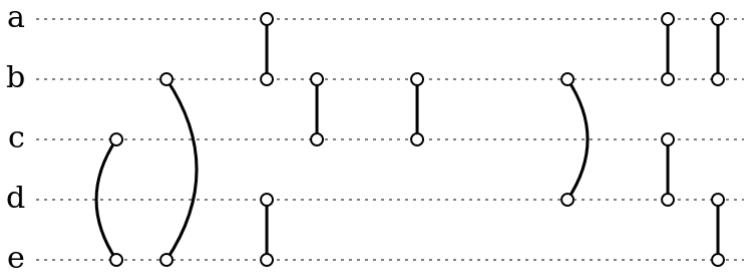


Figure 1: Représentation d'un flot de liens

Un **flot de liens** (schématisé Figure 1) est ici défini comme un triplet (T, V, E) où T est un intervalle de temps, V est un ensemble de nœuds et où E est un ensemble de triplets dans $T \times V \times V$ qui représentent les interactions dans le temps entre les nœuds.

On peut définir un flot de liens **biparti** comme un quadruplet (T, V_1, V_2, E) où T est un intervalle de temps, V_1 et V_2 sont des ensembles de nœuds et où E est un ensemble de triplets dans $T \times V_1 \times V_2$ qui représentent les interactions dans le temps entre les nœuds des deux parties.

De nombreux problèmes concrets peuvent être représentés par cette structure. On peut ainsi rendre compte à la fois des propriétés et caractéristiques structurelles propres aux graphes relationnels (Distance entre deux nœuds, Voisins Communs, Cliques ...), et à la fois des caractéristiques dynamiques des interactions (tendances, saisonnalités ...).

Dans la suite de notre rapport, on définit le cadre suivant :

- T est une période de temps à définir dans l'analyse de notre jeu de donnée dans la partie 3.
- V_1 est un ensemble de clients.
- V_2 est un ensemble de clusters de spectacles dont la construction sera détaillée en sections 4.1 et 4.2.
- E est un ensemble de transactions réalisés entre un client de $V1$, un cluster de $V2$ à un moment t.

Notre objet *flot de liens* combine ainsi l'information structurelle du graphe relationnel du constat des transactions client-cluster et l'information dynamique d'une série temporelle qui retranscrirait les variations au cours du temps de l'activité d'un lien client-cluster.

2.3 Fonction de prédiction

A cet effet, un workflow a été développé dans le language Python basé sur les travaux [1].

Cet algorithme a pour objectif de combiner les données structurelles du graphe ainsi que les données dynamiques du flot, pour évaluer les probabilités de formation de liens sur une période donnée à partir des liens sur la période précédente.

On commence par définir un ensemble de propriétés, qu'on appellera métriques, *a priori* susceptibles de rendre compte d'une tendance des liens à se former. Ces métriques sont à la fois basées sur des propriétés structurelles (Voisinage des nœuds du flot...) et des propriétés dynamiques (Récurrence des liens). On définit une métrique comme une fonction de $V_1 \times V_2$, qui a chaque paire de nœuds (u, v) associe la propension de la paire à être formé au cours de la période d'anticipation.

$$\begin{array}{rccc} f & : & V_1 \times V_2 & \to & \mathbf{R}^+ \\ & & u, v & \mapsto & f(u, v) \end{array}$$

Notre fonction de prédiction sera une combinaison linéaire de ces métriques. Ainsi soient f_1, f_2, ..., f_n, nos n métriques présélectionnées, on aura notre fonction de prédiction à optimiser sous la forme, où les α_i sont les coefficients à optimiser :

$$F(u,v) = \sum_{i=1}^{n} \alpha_i f_i(u,v) \tag{1}$$

2.4 Métriques envisagées dans notre cadre d'étude

Les métriques envisagées dans un premier temps feront le parti pris de la simplicité : extrapolation d'activité par des fonctions linéaires du temps, interactions communes ...

- **Interactions communes asymétriques**
  Soit $u \in V_x$, on définit les voisins de u comme $N(u) = \{w \in V_y | \exists (t) \in T, (t,u,w) \in E\}$

  On définit le terme *interaction*. On dit que $(u,v) \in V_1 \times V_2$ interagissent si $u \in N(v)$ ou de manière équivalente $v \in N(u)$. On dit que $(u,w) \in V_x^2$ interagissent si $\exists v \in V_y \ u \in N(v) \ et \ w \in N(v)$

  Soit $(u,v) \in V_1 \times V_2$ on définit, pour les besoins de nos travaux.

$$IC(u,v) = card(N(v) \bigcap (\bigcup_{w \in N(u)} N(w))) \tag{2}$$

  On notera par ailleurs que cette définition n'est pas symétrique. $IC(u,v) \neq IC(v,u)$. Par convention on choisit $IC(u,v)$ comme étant le nombre d'interactions communes entre u et v dans l'ensemble de provenance de u

- **"Jaccard" asymétrique pour graphes symétriques**
  Soit $(u,v) \in V_1 \times V_2$ on définit, pour les besoins de nos travaux :

$$J(u,v) = \frac{card(N(v) \bigcap (\bigcup_{w \in N(u)} N(w)))}{card(N(v) \bigcup_{w \in N(u)} N(w))} \tag{3}$$

  Par analogie avec la transition Voisins Communs -> Indice de Jaccard, la différence entre notre mesure IC et notre mesure "Jaccard" est la normalisation par rapport à l'ensemble des liens possibles. On a donc $0 \leq J(u,v) \leq 1$ et on vérifie que $J(u,v) = 1$ signifie que les interactions de v dans l'ensemble de provenance de u sont les mêmes que les interactions de u dans son propre ensemble.

- **Extrapolation de l'activité**

  Partant de l'intuition que plus un lien est présent pendant la phase d'observation, plus il est susceptible d'être présent pendant la période d'anticipation. On définit l'activité :

$$A_{u,v} = |\{(x,u,v) \in E\}| \tag{4}$$

- **Extrapolation de l'activité récente**

  Partant de l'intuition que plus un lien a été présent dans une période récente, plus il est susceptible d'être présent pendant la période d'anticipation. Notons δ une durée et Ω la date de fin de période d'observation, on définit :

$$A_{\delta,(u,v)} = |\{(x,u,v) \in [\Omega - \delta, \Omega]\}| \tag{5}$$

  Un autre moyen d'extrapoler l'activité récente et de déterminer la date du k-ième dernier lien formé, on définit donc t_k tel que:

$$t_k = \max(t : |\{(x,u,v) \in E, t \leq x \leq \Omega\}| = k) \tag{6}$$

  Puis :

$$A_{k,(u,v)} = \frac{k}{\Omega - t_k} \tag{7}$$

  Afin d'avoir une activité qui tend vers l'infini quand t_k tend vers Ω

• Optimisation de la fonction de prédiction

L'objectif de cette étape est la maximisation de la précision des liens prédits par l'algorithme. On commence par redéfinir le F-score dans notre cadre d'étude qui sera notre mesure de qualité :

En notant $N_{u,v}$ le nombre de liens prédits (un réel positif) entre u et v et $N'_{u,v}$ le nombre de liens qui ont réellement été formés, on définit par analogie :

$$TP_{u,v} = \min(N'_{u,v}, N_{u,v}) \tag{8}$$

$$FP_{u,v} = \max(N_{u,v} - N'_{u,v}, 0) \tag{9}$$

$$FN_{u,v} = \max(N'_{u,v} - N_{u,v}, 0) \tag{10}$$

Cette analogie est motivée par l'intuition que les vrais positifs sont des liens à la fois prédits et existants, que les faux positifs sont des liens prédits mais inexistants, et les faux négatifs sont des liens non prédits alors qu'ils existaient.

La précision ($\frac{TP}{TP+FP}$), le rappel ($\frac{TP}{TP+FN}$) et le F1-score habituels découlent évidemment de ces définitions ($\frac{2 \cdot precision \cdot recall}{precision+recall}$).

Afin d'optimiser notre algorithme, on définit comme ci dessus (Figure 2) deux périodes de temps : une période d'entraînement (ou d'optimisation de la fonction de prédiction) et une période de test. Chacune de ces deux périodes est subdivisée en deux sous-périodes : une période d'observation et une période d'anticipation. Dans la période d'optimisation, on essaye d'ajuster au mieux les paramètres $\{\alpha_i\}_i$ de la fonction de prédiction.

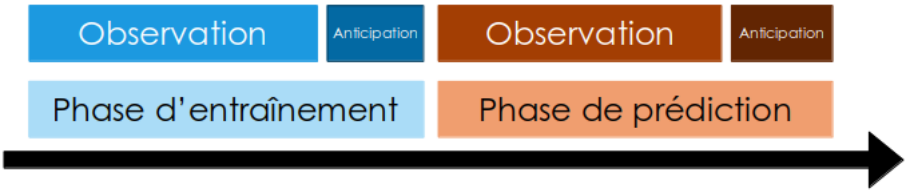


Figure 2: Division temporelle pour l'optimisation de la fonction de prédiction

3
DESCRIPTION DU JEU DE DONNÉES

Cette partie est consacrée à la description du jeu de données sur lequel j'ai travaillé. Elle se livre aussi à quelques observations qui motiveront certains choix faits au cours de la partie 4.

3.1 Structure des données

Les données ont été récupérées et compilées par Delight. Les données sont séparés en trois tables de données distinctes.

- **Shows** (Figure 3) : Contient des informations que l'on qualifiera de "sémantiques" relatives aux spectacles : descriptions texte, "tags", noms, identifiants unique ... On compte plus de 83 000 shows distincts dans ce dataset.

- **Events** (Figure 4) : Contient des informations relatives à la représentation même d'un spectacle : ville, salle ou lieu de l'évènement, date, identifiant du spectacle représenté, identifiant unique de l'évènement. Ainsi plusieurs évènements peuvent être une représentation d'un même spectacle à des dates/lieux différents, l'inverse n'étant pas vrai, chaque évènement n'étant la représentation que d'un seul spectacle. Il y a plus de 693 000 évènements recensés.

- **Transactions** (Figure 5) : Contient des informations relatives aux transactions effectuées sur les billetteries ayant partagé leurs données, identifiants de l'évènement et du spectacle, quantité de tickets achetés, catégorie du ou des ticket(s) acheté(s), informations relatives à la transaction (date, billetterie), informations relatives au client (identifiant unique, ville, département, genre). On compte plus de 16 millions de transactions réalisées par 1.5 million d'utilisateurs uniques.

Les données sont d'une qualité hétérogène. Les datasets sont issus d'un agrégat de différentes billetteries ayant partagé leurs données avec Delight. Les données sont donc sur certains champs assez lacunaires, voire très lacunaires (Figure 6a). Un travail d'harmonisation est réalisé ces derniers temps chez Delight pour mieux unifier ces données billetteries.

On peut également donner une idée de la composition du jeu de données par l'observation de la proportion de *tags* (Figure 6b courants dans le champs *event provider types*)

	event_stakeholders	event_provider_types	event_name	event_provider_description	delight_show_id
68205	["Astonvilla"]	["concert","variete internationale"]	ASTONVILLA BULLITT	Astonvilla, dont le son rock s'est imposé au f...	f859c73c-1f6b-42ff-8e15-2ed79d0fa7df
21478	["Georges Bizet","Yorgos Loukos"]	["danse classique"]	CARMEN / L'ARLESIENNE	Avant de prendre la direction du Ballet de l'O...	12cb086f-fcac-401e-87ce-5466f9b0c744
22636	["Jeanne Balibar","Jean-Damien Barbin","Mikhai...	["théâtre contemporain"]	DAS LEBEN DES HERRN DE MOLIERE	Textes Mikhaïl Boulgakov, Pierre Corne...	61178034-8081-46fb-855a-3ec83b977519
25347	NaN	["activités de loisirs divers"]	COURS PHOTO INITIATION 3H MARSEILLE	<p>\tL'objectif est de quitter le mode automat...	668c16c4-e6c3-4fe0-a321-5c73b2d67b12
78337	["Bear's Den"]	["pop-rock/folk"]	BEAR'S DEN	<p align="justify"><font color="#ff0000"...	09b600ff-6d60-49ef-b5f7-7fef716cd629

Figure 3: Aperçu de la table de données "Shows"

	event_datetime	venue_city	event_name	performance_provider_id	venue_name	delight_ige_id	delight_event_id	delight_show_id
666843	2017-07-08T21:30:00.000Z	PARIS 3	TRISTAN LOPIN DANS DEPENDANCE AFFECTIVE	5530578.0	Théâtre du Marais	a7882dce-0252-4442-b3e8-037ccfd662e6	73fece20-0159-475c-8c4b-7b06726440bc	e769e29b-0d0b-469d-b0fb-29f43942c9d6
116248	2016-10-19T21:00:00.000Z	CAHORS	LA YEGROS	NaN	LES DOCKS	c50a8ce6-3f42-4bec-b70b-608b5248baaa	72335098-d1ef-4655-b4e2-eacfae62bd35	fd68b3fd-c02a-4aff-9cbf-97e0fb98059d
109035	2016-10-08T16:00:00.000Z	LE KREMLIN BICETRE	LA PETITE CASSEROLE D ANATOLE	NaN	ESPACE CULTUREL ANDRE MALRAUX	95b4ee3b-398b-47ba-a3a8-f238de9f015c	fedc0c14-e5be-4c0a-8a76-91368494f1f8	56615262-4ec2-490f-bdd5-86e4d0ea1ab1
692079	2016-11-13T17:00:00.000Z	PARIS 11	YVETTE GUILBERT IL NE FAUT JAMAIS SE DECOURAGER	5103841.0	Comédie Nation	bb016bec-42f7-46d4-8cb5-7f60b2d47e35	2b2e6abe-7813-4070-835c-29a66e9e3d69	1b00f7f4-2df7-4239-b1dd-08926b4ce256
535488						3c697afa-		

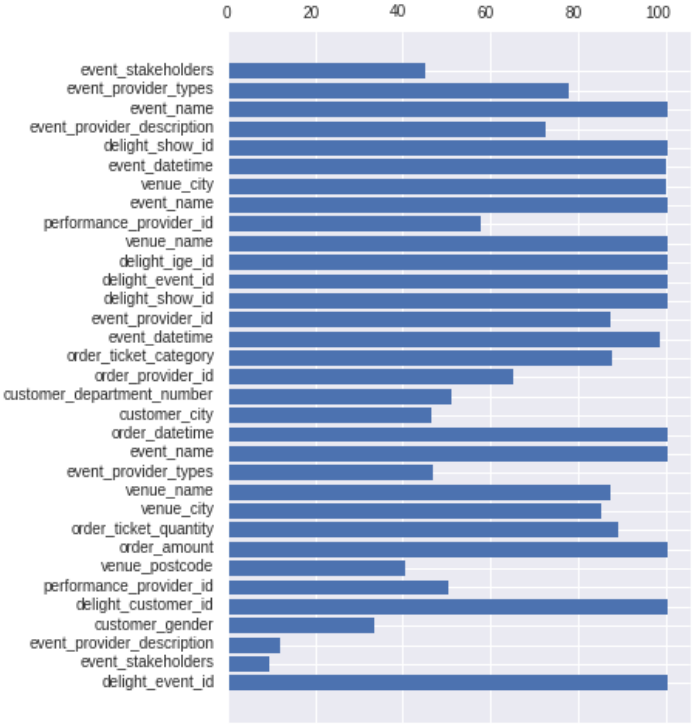
Figure 4: Aperçu de la table de données "Events"

	event_provider_id	event_datetime	order_ticket_category	order_provider_id	customer_department_number	customer_city	order_datetime	event_name	ev
1217063	NaN	2015-02-14 20:30:00	Catégorie 2	8788	94.0	IVRY SUR SEINE	2015-02-06 13:03:32.427	KAMEL LE MAGICIEN	
693775	304096	2014-11-09 20:30:00	NaN	104121188	NaN	NaN	2014-09-20 10:39:18.000	KRAFTWERK 4 THE MAN MACHINE	
1038101	NaN	2016-01-22 20:00:00	Catégorie 3	189869.0	94	NaN	2015-09-09 19:24:12.330	HOZIER	
372340	139568	2015-08-29 20:30:00	Catégorie unique	NaN	75	PARIS 16	2015-08-28 18:32:00.000	ARLEQUIN VALET DE DEUX MAITRES	
46959	PADEC	2017-02-22 10:30:00	Catégorie unique	GSE1073256832893	NaN	NaN	2017-01-24 17:03:20.000	EXPOSITIONS PALAIS DE LA DECOUVERTE	

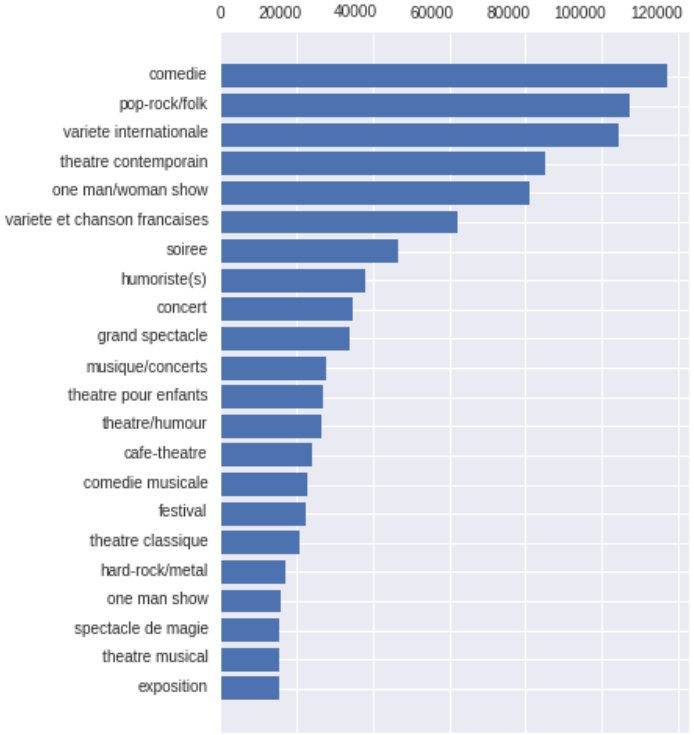
Figure 5: Aperçu de la table de données "Transactions"

3.2 Analyse des données temporelles du jeu de données et définitions

Commençons par nous livrer à une exploration du jeu de données afin de fournir des pistes d'études à une meilleure prise en compte du profil temporel des évènements.



(a) Pourcentage de remplissage de chacun des champs dans le dataset



(b) Nombre de transactions relié aux tags les plus courants (sur un échantillon aléatoire d' 1,5 million de transactions

Notre analyse préliminaire sera de chercher à identifier des points de discrimination sur lesquels travailler. Pour cela en nous basant sur une classification des évènements primaires : les "tags" des spectacles attribués sous la responsabilité des billetteries. On espère voir apparaître des différenciations pertinentes selon ces tags.

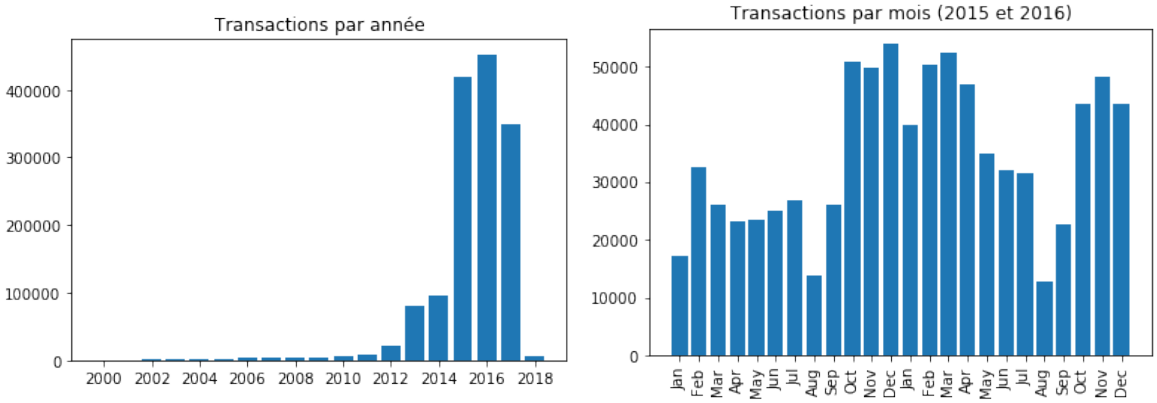


Figure 7: Répartition des transactions sur différentes périodes

• Analyse Générale de la composition du dataset

L'essentiel des données disponibles, se concentrent sur les années 2015, 2016, 2017 (Figure 7). Il nous est déjà possible de voir une évolution du volume de transactions réalisées au cours de l'année avec une forte baisse des transactions au cours du mois d'Août par exemple, baisse plus marquée sur certains tags que sur d'autres.

• Analyse relative à l'anticipation

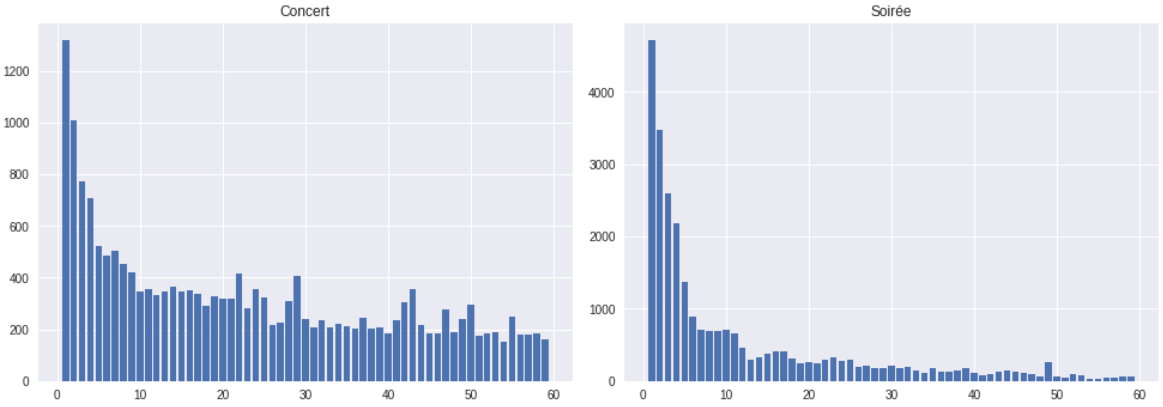


Figure 8: Profil de l'anticipation agrégée pour les tags "Soirée" et "Concert"

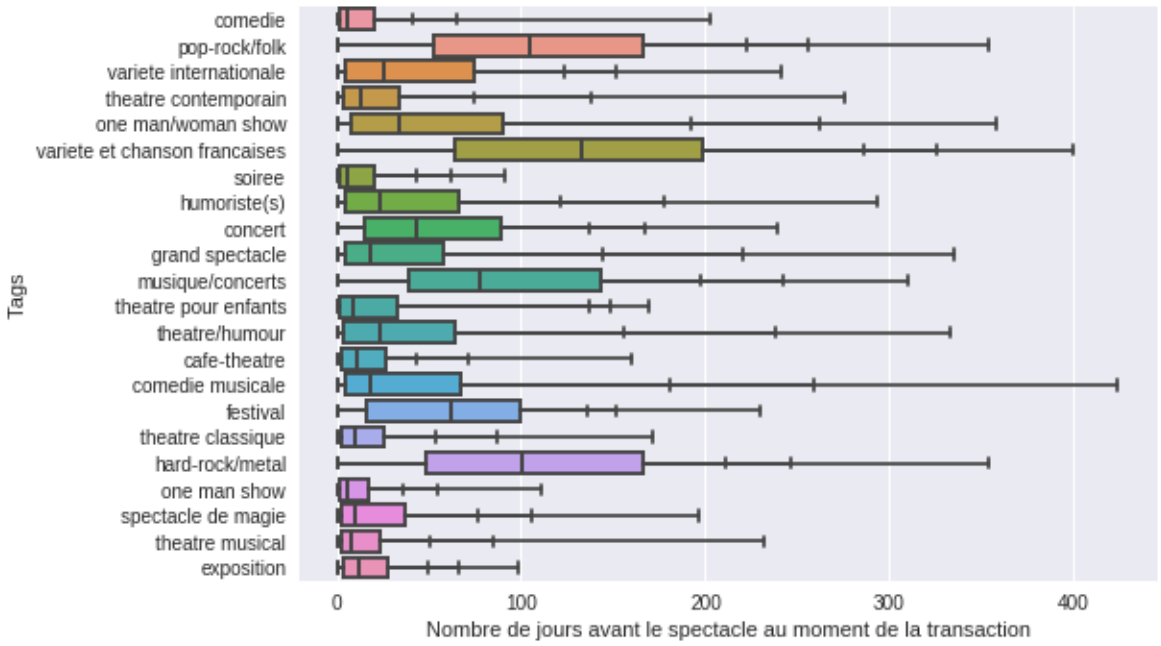


Figure 9: Profil de l'anticipation pour différents tags
Boîte 1er quartile, médiane, 3e quartile
Traits Dernier décile, vingtile, centile

On choisit de définir par le terme **anticipation**, la donnée du temps écoulé entre la réalisation d'une transaction et le moment de l'évènement. Elle quantifie la tendance d'un client à anticiper un évènement. Cette donnée est apparue comme importante en raison des différences de profil d'un type d'évènement à l'autre (Figure 8)

De cette donnée on peut étudier en fonction du tag la tendance moyenne des clients. On voit tout de suite apparaître des différences de profil importantes d'un type de spectacle à un autre. Ainsi, en représentant les diagrammes en boîte relatifs à cette donnée (Figure 9) pour chaque tag, on voit apparaître des comportements très différents selon que le spectacle est identifié comme « variété et chansons françaises » (souvent associé à de grands concerts dans de grandes salles avec des artistes populaires) ou « soirée ».

• Analyse relative au calendrier

On désigne par **calendrier**, la répartition des volumes d'achats au cours d'une semaine, jour par jour, tranche horaire par tranche horaire. L'objectif étant de repérer des pics d'activités propres à certains types d'évènements.

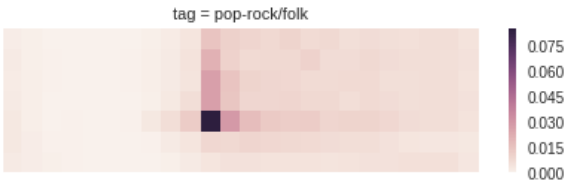
Cette donnée est difficilement exploitable en raison des habitudes d'horaires de mise en vente de billets de la part de certains distributeurs de billets qui débouchent à cer-

taines tranches horaires sur des pics d'activités "artificielles". Ceci est particulièrement visible pour le tag *pop-rock/folk* (Figure 10c), où les billets des gros concerts catégorisés par ce tag partent en quelques minutes après la mise en vente des billets (souvent à 10h le vendredi).

Il est possible d'identifier des différences de comportement entre les différentes catégories de spectacles. Ainsi l'activité relative au tag *festival*, relative donc à des évènements de plusieurs jours, souvent assez chers est plus importante dans l'après-midi en semaine ce qui peut suggérer des décisions d'achats plus réfléchies, plus anticipées. Le tag *soirée* à l'inverse relatif à des évènements nocturnes festifs connaît un pic d'activité les vendredis et samedis soirs, ce qui suggère des achats impulsifs et des billets pris aux derniers moments.

(a)

(b)



(c)

Figure 10: Répartition de l'activité des transactions pour différents tags au cours de la semaine (heures de 0 à 24 en abscisses, jours de Lundi à Dimanche en ordonnée)

4
APPROCHE POUR LA RECOMMANDATION D'ÉVÈNEMENTS À DES CLIENTS

Dans cette partie, nous utilisons les outils décrits précédemment pour mettre en place une approche d'un système de recommandation de spectacles à des clients. Nous avons choisi de décomposer notre système en plusieurs étapes :

- **Pré-traitement** : développée dans la section 4.1, il vise à la vectorisation des informations sémantiques et des informations temporelles des spectacles afin de préparer la clusterisation
- **Clustering** : développée dans la section 4.2, elle vise au regroupement de spectacles en clusters. Ce regroupement est nécessaire pour créer une récurrence de liens. On cherchera à partir de là à étudier l'activité des liens client-cluster
- **Prédictions de Liens** : après construction du flot de liens immédiatement après l'étape de clustering. On cherchera dans la section 4.3 à trouver le bon paramétrage (périodes d'apprentissage, choix de métriques) pour obtenir des prédictions les plus pertinentes possibles.
- **Recommandation** : Une fois nos prédictions de liens client-cluster établis on extrait de ces clusters des spectacles compatibles avec le client (bon timing, bonne localisation géographique...)

Figure 11: Protocole de recommandation

4.1 Exploitation des données temporelles et sémantiques pour la clusterisation des évènements

Cet axe de travail a pour but de réaliser des clusters d'évènements similaires en prenant en compte à la fois les caractéristiques sémantiques et les caractéristiques temporelles d'un évènement..

• Vectorisation d'un évènement

Cette première étape vise à réaliser une synthèse des différentes informations dont l'on dispose sur un évènement, on compte :

- Les informations sémantiques : noms, tags, artistes ...
- Les informations temporels : anticipation, calendrier

• Vectorisation des données sémantiques

Les données sémantiques sont regroupés dans la table de données *shows* relatives donc au contenu même des spectacles. Elles sont regroupés en 4 champs : name, stakeholders, types (*tags*), description.

Comme le traitement des données sémantiques n'est pas un objectif central de notre stage on décide de se focaliser sur les données *tags* qui nécessitent un traitement moins important. Notre objectif est donc de trouver un moyen de convertir ces tags en vecteurs numériques, permettant par la suite de convertir les évènements en vecteurs numériques également. Le protocole de vectorisation des tags sera le suivant :

- Établissement d'une mesure de similarité entre les tags, basée sur la co-occurrence des tags entre eux.
- Construire la matrice des similitudes $(s(x_i, x_j))_{i,j}$, selon les mesures décrites section 2.1, puis appliquer un algorithme de réduction de dimensions (PCA, Isomap ...). Les vecteurs issues de ces algorithmes constitueront la partie "sémantique" du vecteur numérique lié au spectacle.

• Vectorisation des données temporelles

L'observation des données temporelles (se référer à la partie 3) nous a donné l'occasion d'observer deux informations temporelles importantes. Premièrement, on considère la série temporelle des achats par rapport à la distance au jour de l'évènement, à savoir l'anticipation. Deuxièmement, on considère la distribution temporelle des achats au cours d'une période définie à savoir le calendrier.

- **Anticipation** : L'idée d'exploiter l'anticipation pour la vectorisation résulte de l'intuition que des comportements clients différents, sont reliés à des publics différents, donc à des segments marchés différents. On décide d'adopter une démarche consistant à définir des paliers d'anticipations : anticipation de moins d'un jour, de 1 à 3 jours, de 3 à 7 jours ... Et pour chaque évènement de rendre compte de la distribution des transactions dans les différentes classes. On espère ainsi rendre compte au mieux du profil temporel d'un évènement en dimension réduite (plusieurs profils testé à 10 ou 20 paliers ...).

  D'un point de vue plus formel on définit une suite de paliers $p_0 = 0, p_1, p_2, ..., p_n$, et on définit la variable $AnticipationDelta_i = \frac{billets\ achetés\ sur\ la\ période\ [p_{i-1}, p_i]}{billets\ achetés}$

- **Calendrier** : De la même façon des comportements clients différents peuvent donner lieu à des segments marchés différents. On décide de fractionner la semaine en jours et tranches horaires (0-3h, 3-6h , ...), puis comme pour l'anticipation de rendre compte de la distribution de l'évènement dans ce vecteur. On définit la encore une suite de paliers $p_0 = Lundi\ Minuit, p_1, p_2, ..., p_n = Dimanche\ 23h59$ et on définit la variable $Calendrier_i = \frac{billets\ achetés\ sur\ la\ période\ [p_{i-1}, p_i]}{billets\ achetés}$

4.2 Clustering d'évènements

Cette étape a pour but de séparer les évènements en groupes d'évènements similaires. Un clustering naturel serait un clustering basé sur les tags. Nous avons cependant pour objectif de raffiner un peu ce clustering en y ajoutant les données temporelles d'anticipation et de calendrier précédemment détaillées.

Pour cela on considère pour chaque évènement le vecteur numérique résultant de la concaténation des vectorisations effectuées lors de la première étape.

On réalise par la suite un clustering de ces vecteurs à l'aide de techniques classiques de vectorisation : k-moyennes, k-moyennes + clustering hiérarchique ... Afin d'avoir des clusters plus équilibrés on appliquera également des techniques de type k-moyennes équilibrés (Chang et al, 2014 [2])..

Le nombre de clusters est évidemment très modulables 10, 100, 500 clusters, il est pour l'instant difficile de savoir quel ordre de grandeur on vise. Sachant que des grands clusters peu nombreux augmentent le volume de transactions disponibles pour les liens clients-cluster mais regroupent des évènements qui n'ont rien à voir entre eux. Il y a donc un arbitrage à réaliser entre la nécessité d'avoir des clusters d'évènements suffisamment similaires et la nécessité d'avoir un volume de transactions par cluster suffisant.

4.3 Prédiction de liens client-cluster

• Construction du flot de liens

Une fois ce clustering réalisé on vise à construire un flot de lien exploitable. Ce flot de lien reliera des nœuds représentant de utilisateurs à des nœuds représentant des clusters d'évènements.

Pour les besoins de l'algorithme, nous construisons donc des flots de liens de la forme $t\ u\ c$, où t est le temps écoulé en minutes depuis le 1er Janvier 2016 au moment de la transaction. u est un identifiant unique pour l'utilisateur et c le cluster d'évènement avec lequel l'interaction se produit.

Pour nos premiers tests, nous générerons des flots de lien réduits : un échantillon aléatoire de 5000 clients ayant réalisés entre 20 et 100 transactions chacun entre le 1er Janvier 2016 et le 31 Décembre 2017. Ces échantillons ont vocations à être étendus dans la suite de nos travaux.

• Exploitation du flot de liens

Le flot de liens est exploité dans un objectif de prédiction de liens futurs par une réadaptation de l'algorithme [1].

Les tests seront effectués avec une large palette de flot de liens :

- Construits à partir de clusterisations issues de vectorisation différentes (section 4.1)
  - Dimension de la vectorisation des tags
  - Présence ou non des données d'anticipation et de calendrier ...
- Construits à partir de de clusterisations réalisés par des méthodes différentes (section 4.2) : k-moyennes, balanced k-moyennes
- Construits à partir de clusterisations plus ou moins larges : nombres de clusters

Ces tests seront également effectué à partir de configuration de l'algorithme différentes :

- Observation sur 6 mois, Prédiction sur 6 mois (Six-Six). Observation sur 9 mois, Prédiction sur 3 mois (Nine-Three) ...
- Utilisation de métriques différentes parmi celles proposées : Extrapolation, Extrapolation du dernier mois, Extrapolation des 3 derniers mois, Extrapolation des 5, 10, 20 derniers liens. Pour les liens récurrents. Voisins Communs, Indice de Jaccard, Similitude des Clusters (...) pour les nouveaux liens.

4.4 Application à la recommandation

• Premiers Résultats

Pour nos premiers tests, nous avons tenté de trouvé des configurations de l'algorithme qui donnait des résultats convenables. Nous détaillons ici nos propositions pour les différents axes de configuration de l'algorithme

- La **vectorisation** dans un but de clusterisation. La base de notre vectorisation a été de se baser sur les Tags. Nous avons ensuite essayé d'ajouter aux vecteurs d'évènements les données d'Anticipation (Tags + Ant) et de Calendrier (Tags + Ant + Cal).
- Le **nombre de clusters** d'évènements : de 10 à 100
- La **paramétrisation** de l'algorithme : Deux modes principaux ont été choisit.
  - Dans un premier mode on choisit d'entraîner sur une année et de prédire sur l'autre. Par exemple pour une paramétrisation Six-Six, on choisit d'apprendre sur l'année 2016 et de prédire sur l'année 2017, pour chaque année, on observe les 6 premiers mois pour anticiper les 6 derniers (Parallèlement pour Onze-Un et Neuf-Trois).
  - Dans un second mode, le mode **glissant**, on fait "glisser" la fenêtre d'observation et d'anticipation. Ainsi pour une paramétrisation Neuf-Trois Glissant, on apprend sur période Janvier-Décembre 2016 (9 mois d'observation, 3 mois d'anticipation) et on prédit sur la période Mars 2016-Février 2017. Cela à l'avantage de réduire l'éloignement entre l'entraînement et la prédiction mais le désavantage d'effacer des observables saisonnalités annuelles.

Nous mettons ci-dessous quelques résultats significatifs obtenus. Pour quantifier nos résultats, nous avons choisi les redéfinitions naturelles de la prédiction, du recall, du F1-score plus haut. On met en lumière ci-dessous les résultats significatifs déjà obtenus.

Vectorisation	**Clusters**	**Paramétrisation**	**F1-New**	**F1-Few**	**F1-Recurrent**
Tags + Ant + Cal	100	Six-Six	0.08	0.45	**0.63**
Tags + Ant + Cal	20	Six-Six	0.05	0.48	0.57
Tags + Ant + Cal	10	Six-Six	0.05	0.52	0.57
Tags + Ant + Cal	20	Neuf-Trois	0.02	0.29	**0.45**
Tags + Ant + Cal	10	Neuf-Trois	0.02	0.32	0.42
Tags + Ant + Cal	20	Neuf-Trois Glissant	0.015	0.34	**0.39**
Tags + Ant	20	18-Trois Glissant	0.08	0.22	0.27
Tags + Ant + Cal	10	Onze-Un	-1	0.14	0.19
Aleatoire	20	Six-Six	0.12	0.17	-1

5 DÉVELOPPEMENTS FUTURS

5.1 Amélioration de la qualité du clustering

L'une des faiblesses identifiées est la qualité assez faible du clustering actuelle, dans le sens. En effet :

- Les données sémantiques sont souvent assez lacunaires ce qui empêche de les exploiter (voir tableau ci-dessous) à un but de clustering.
- Les vectorisations actuelles prenant en compte les données temporelles n'ont pas donné lieu à une amélioration réelle du clustering.
- Les methodes de clustering actuelles ont soit donné lieu à un clustering trop déséquilibré en terme de volume, soit à des temps de calculs trop longs.

Les solutions envisagées actuellement passeront donc essentiellement par une meilleure prise en compte des données sémantiques.

Il est envisagé d'utiliser pleinement les autres données sémantiques, à savoir les stakeholders (troupes, artistes impliqués...), les noms et descriptions des évènements. Elles répondent à des problématiques différentes car elles ne forment pas une classification aussi immédiate que les tags : le vocabulaire est infiniment plus vaste, présente de nombreux "déchets" et doit donc faire l'objet d'un travail de processing à part. De plus, ces champs sont plus lacunaires et réduiraient donc notre volume de spectacles exploitables. Le tableau suivant représente la diminution progressive du nombre de spectacles disponibles en ajoutant progressivement les différents champs sémantiques.

Champs ajouté	**Intersection**
event name	83034
event provider type (*tags*)	64335
event stakeholders (artistes ...)	37189
event description (texte/sinopsis)	35779

5.2 Utilisation de fonctions et algorithmes plus complexes

Nous avons pour l'instant utilisé une fonction très simple qui est une combinaison linéaire de métriques pour nos prédictions. Nous avons pour ambition d'essayer d'améliorer nos prédictions par l'utilisation d'outils plus complexes :

- **Kernelisation** (polynomiales, exponentielles, ...) des métriques afin de permettre de ressortir des variables plus pertinentes avant regression.
- Utilisation de régresseurs plus complexes. L'objectif étant de transformer les métriques dynamiques et structurelles en variables pour des algorithmes avancés (RandomForest, GradientBoosting, RidgeRegression ...).
- Création de classes de liens : pour l'instant les liens ont été séparés en 3 classes : nouveau (0 occurrence), peu récurrents (< 5 occurrences), récurrents. Chaque classe faisant l'objet d'une optimisation des coefficients (comme expliqué section 2.2) à part. L'objectif serait de changer la conception de ces classes de liens.

Pour cela il est envisagé d'utiliser d'autres mesures de qualités que le f1-score et d'adapter la méthodologie employée jusque maintenant. L'objectif sera d'extraire proportionnellement dans les clusters recommandés selon une distribution anticipée par nos algorithmes de prédictions. Il sera important, plus tard, dans un objectif d'optimisation des nombreux hyper-paramètres possibles (vectorisations, clustering, régression ...) d'avoir une fonction objective afin de comparer du mieux possible les paramétrages.

Nous avons jusque maintenant comparé l'activité prédite et l'activité réelle dans nos flots de lien. Pour évaluer nos modèles, on cherchera plus tard à comparer plutôt que l'activité réelle et l'activité prédite la distribution dans les faits et la distribution réelle pour chaque client de l'échantillon. On note $d_e(u, c)$ la distribution réelle sur la période de prédiction qu'on cherche à prévoir. $d_e(u, c)$ est donc la proportion d'évènement du cluster c parmi l'ensemble des évènements auxquels a pris par le client u sur la période cible. En particulier :

$$\forall u \sum_c d_e(u, c) = 1 \tag{11}$$

De même la distribution prédite sera noté d_p.

On cherche pour chaque client u à approximer au mieux la fonction $d_e(u, .)$ par la fonction $d_p(u, .)$. On propose donc de minimiser le nombre suivant pour chaque u, afin de se rapprocher d'une fonction d'erreur de type **erreur quadratique moyenne**

$$J_u = \sum_c (d_p(u, c) - d_e(u, c))^2 \tag{12}$$

Sur l'échantillon entier on cherchera donc à minimiser l'un des nombres suivants :

$$J = \frac{1}{n} \sum_u J_u = \frac{1}{n} \sum_u \sum_c (d_p(u, c) - d_e(u, c))^2 \tag{13}$$

$$J = \frac{1}{n} \sum_u J_u^2 = \frac{1}{n} \sum_u (\sum_c (d_p(u, c) - d_e(u, c))^2)^2 \tag{14}$$

6 CONCLUSION

Finalement, on constate que l'approche utilisée est prometteuse. Les outils utilisés sont pour l'instant très simples (Utilisation de fonctions constantes, de métriques simples, de jeux de données réduits) et les vaste pistes de développements que nous avons évoqués laissent entrevoir la possibilité de construire des moteurs de recommandations de bonne qualité à l'aide du formalisme des flots de liens. L'utilisation à tous les niveaux de modèles plus fins dans la seconde partie de mon stage permettra sans doute d'avoir une meilleure idée des possibilités de développement d'un tel système de recommandation.

D'un point de vue plus personnel, ce stage me permet de travailler avec un nouvel outil et un nouveau formalisme de haut niveau et à fort potentiel et de l'appliquer à un problème concret. Il s'agit d'une expérience de recherche qui sera riche en enseignement pour la suite de ma formation. De plus, j'ai pu travailler de nombreux points techniques qui s'inscrivent directement dans ma formation : manipulation et visualisation de données, séries temporelles, graphes, apprentissage supervisé ou non supervisé, réduction de dimensions...

Je remercie de fait celles et ceux qui permettent le bon déroulement de mon stage, en premier lieu Lionel Tabourier, qui m'a proposé ce stage et me suit tout au long de celui-ci. Également Jean Creusefond, pour m'avoir également suivi et conseillé au cours de mon stage. Je remercie également toute l'équipe de Delight pour son accueil et pour avoir mis à ma disposition les ressources et les données nécessaires à la réalisation de ce stage.

REFERENCES

[1] Thibaud Arnoux, Lionel Tabourier, and Matthieu Latapy. Predicting interactions between individuals with structural and dynamical information. *arXiv:1804.01465 [physics]*, March 2018. arXiv: 1804.01465.

[2] Xiaojun Chang, Feiping Nie, Zhigang Ma, and Yi Yang. Balanced k-Means and Min-Cut Clustering. *arXiv:1411.6235 [cs]*, November 2014. arXiv: 1411.6235.

[3] Jean Creusefond and Matthieu Latapy. Propagation of content similarity through a collaborative network for live show recommendation. *arXiv:1804.09073 [cs]*, April 2018. arXiv: 1804.09073.

[4] Souvik Debnath, Niloy Ganguly, and Pabitra Mitra. Feature weighting in content based recommendation system using social network analysis. page 1041, 2008.

[5] Matthieu Latapy, Tiphaine Viard, and Clémence Magnien. Stream Graphs and Link Streams for the Modeling of Interactions over Time. December 2017.

[6] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based Recommender Systems: State of the Art and Trends. pages 73–105, 2011.

[7] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender Systems: Introduction and Challenges. pages 1–34, 2015.

[8] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based Collaborative Filtering Recommendation Algorithms. pages 285–295, 2001.